Development and prospective application of chemoinformatic tools to explore new ligand chemistry and protein biology

Dissertation

zur Erlangung des Doktorgrades
der Naturwissenschaften
(Dr. rer. nat.)

dem
Fachbereich Pharmazie der
Philipps-Universität Marburg
vorgelegt von

Jakub Gunera, M. Sc.

aus

Słubice

Marburg/Lahn 2017

Erstgutachter

Prof. Dr. Peter Kolb
Institut für Pharmazeutische Chemie
Philipps-Universität Marburg

Zweitgutachter

Prof. Dr. Shu-Ming Li
Institut für Pharmazeutische Biologie und Biotechnologie
Philipps-Universität Marburg

Eingereicht am 19.05.2017

Tag der mündlichen Prufüng am 10.07.2017

Hochschulkennziffer: 1180

dla mamy, siostry i brata

Omnis mundi creatura
quasi liber et pictura
nobis est, et speculum.
Nostrae vitae, nostrae mortis,
nostri status, nostrae sortis
fidele signaculum

from **Omnis mundi creatura** by **Alanus ab Insulis**, † 1220.

Abstract

Drug discovery and design is a tedious and expensive process whose small chances of success necessitates the development of novel chemoinformatic approaches and concepts. Their common goal is the efficient and robust identification of promising chemical matter and the reliable prediction of its properties. Computer-aided drug discovery and design (CADDD) and its multifarious installments throughout the different phases of the drug discovery pipeline contribute significantly to the expansion of the hits, the understanding of their structure-activity relationship and their rational diversification. They alleviate the development's costs and its time-demand thus support the search for the needle in the haystack – a potent hit. The HTS-driven brute-force nature of current and of the decades' past discovery and design strategies compelled researchers to develop ideas and algorithms in order to interfere with the pipeline and prevent its frequent failures. In the introduction, I describe the drug discovery and design pipeline and point out interfaces where CADDD contributes to its success.

In **Part 1** of this thesis, I present a novel methodology that supports the early-stage hit discovery processes through a fragment-based reduced graph similarity approach (**RedFrag**). It is a chimeric algorithm that combines fingerprint-based similarity calculation with scaffold-hopping-enabling graph isomorphism. We thoroughly investigated its performance retro- and prospectively. It uses a new type of reduced graph that does not suffer from information loss during its construction and bypasses the necessity of feature definitions. Built upon chemical epitopes resulting from molecule fragmentation, the reduced graph embodies physico-chemical and 2D-structural properties of a molecule. Reduced graphs are compared with a continuous-similarity-distance-driven maximal common subgraph algorithm, which calculates similarity at the fragmental and topological levels.

The second chapter, **Part 2**, is dedicated to **PrenDB**: A digital compendium of the reaction space of prenyltransferases of the dimethylallyltryptophan synthase (DMATS) superfamily. Their catalytical transformations represent a major skeletal diversification step in the biosynthesis of secondary metabolites including the indole alkaloids. DMATS enzymes thus contribute significantly to the biological and pharmacological diversity of small molecule metabolites. The attachment of the prenyl donor to lead- or drug-like molecules renders the prenyltransferases useful in the access of chemical space that is difficult to reach by conventional synthesis. In PrenDB, we collected the substrates, enzymes and products. We then used a newly developed algorithm based on molecular fragmentation to automatically extract reactive chemical epitopes. The analysis of the collected data sheds light on the thus far explored substrate space of DMATS enzymes. We supplemented the browsable database with algorithmic prediction routines in order to assess the prenylability of novel compounds and did so for a set of 38 molecules.

In a case study, **Part 3**, we investigated the regioselectivity of five prenyltransferases in the presence of unnatural prenyl donors. Detailed biochemical investigations revealed the acceptance of these dimethylallyl pyrophosphate (DMAPP) analogs by all tested enzymes with different relative activities and regioselectivities. In order to understand the activity profiles and their differences on a molecular level we investigated the interaction within the enzyme-prenyl donor-substrate system with molecular dynamics. Our experiments show that the reactivity of a prenyl donor strongly correlates with the distance of its electrophilic, reactive atom and the nucleophilic center of the substrate molecule. It renders the first step towards a better mechanistic understanding of the reactivity of prenyltransferases and expands significantly the potential usage and rational design of tryptophan prenylating enzymes as biocatalysts for Friedel–Crafts alkylation.

Lastly, in **Part 4**, we present the synergistic potential of combined ligand- and structure-based drug discovery methodologies applied to the β_2-adrenergic receptor (β_2AR). The β_2AR is a G protein-coupled

receptor (GPCR) and a well-explored target. By the joint application of fingerprint-based similarity, substructure-based searches and docking we discovered 13 ligands – ten of which were novel – of this particular GPCR. Of note, two of the molecules used as starting points for the similarity and substructure searches distinguish themselves from other β_2AR antagonists by their unique scaffold. Thus, the usage of a multistep hierarchical or parallel screening approach enabled us to use these unique structural features and discover novel chemical matter beyond the bounds of the ligand space known so far and emphasize the intrinsic complementarity of ligand- and structure-based approaches. The molecules described in this work allow us to explore the ligand space around the previously reported molecules in greater detail, leading to insights into their structure-activity relationship. In addition, we also characterized our hits with experimental binding and selectivity data and discussed it based on their putative binding modes derived by docking.

Zusammenfassung

Die Wirkstoffentwicklung ist ein mühsamer und teurer Prozess, dessen geringe Erfolgsaussichten die Entwicklung von neuartigen chemoinformatischen Ansätzen und Konzepten erfordern. Ihr gemeinsames Ziel ist die effiziente Identifizierung von vielversprechender chemischer Materie and die verlässliche Vorhersage ihrer Eigenschaften. Die computergestützte Wikstoffentwicklung und ihre vielseitigen Einsatzmöglichkeiten innerhalb der verschiedenen Phasen des Entwicklungsprozesses eines Wirkstoffs trägt signifikant zur Erweiterung der bekannten Hits, zum Verständnis ihrer Struktur-Aktivitäts-Beziehung und zu ihrer rationalen Diversifizierung bei. Sie senkt die Entwicklungskosten, verkürzt die Entwicklungszeit und unterstützt dabei die Suche nach der Nadel im Heuhaufen – einem potenten Hit. Die von dem Hochdurchsatz-Screening bestimmten Strategien der heutigen und ehemaligen Wirkstoffentwicklungskampagnen machen die Entwicklung von Konzepten und Algorithmen erforderlich, die mit dem Wikstoffentwicklungsprozess inteferieren und sein häufiges Scheitern verhindern. In der Einleitung beschreibe ich diesen Prozess und führe an, an welchen Stellen und wie computergestützte Methoden zum Erfolg einer Kampagne beitragen können.

Im ersten Teil dieser Arbeit stelle ich eine neue Methode vor (**RedFrag**), die in der frühen Phase der Wirkstoffentwicklung zur Entdeckung weiterer Hits eingesetzt werden kann. Dabei handelt es sich um einen chimären Algorithmus, der fragmentbasierte Abstraktion oder Reduktion von Molekülen und eine auf Fingerprints basierende Ähnlichkeit mit graphtheoretischen Konzepten kombiniert. Dadurch wird die Entdeckung neuer chemischer Grundgerüste (Scaffolds) ermöglicht. RedFrag übersetzt Moleküle in ihre reduzierte Form, die im Gegensatz zu verwandten Methoden den Informationsverlust während der Reduktion gering hält und darüberhinaus ohne die aufwendige und rigide Kodierung von chemischen Epitopen, Substrukturen, auskommt. Die fragmentbasierte Reduktion der Moleküle stellt sicher, dass physikochemische Eigenschaften der Fragmente in Form von Fingerprints erhalten bleiben und die relative zweidimensionale Orientierung der Fragmente zueinander in die Berechnung der Ähnlichkeit einbezogen wird.

Der zweite Teil beschäftigt sich mit **PrenDB**. Dabei handelt es sich um eine Zusammenstellung des Reaktionsraumes von Prenyltransferasen aus der Familie der Dimethylallylsynthasen (DMATS). Die Prenyltransferasen katalysieren die Anknüpfung eines Prenylrestes an kleine Moleküle und stellen dabei einen wichtigen Diversifizierungsschritt des molekularen Gerüsts dar. Dadurch spielen sie eine elementare Rolle in der Biosynthese einer Vielfalt von sekundären Metaboliten, einschließlich von Indolalkaloiden. Der katalytische Transfer von Prenylresten an wirkstoffähnliche Moleküle verleiht den Prenyltransferasen eine besondere Signifikanz in Hinblick auf ihren Einsatz in der Diversifizierung von Molekülen und dem Zugang zur neuartiger Chemie, jenseits des Kanons der medizinal-chemischen Synthese. In PrenDB habe ich die Substrate, Produkte, Cofaktoren (Prenylrest-Donatoren) und Enzyme zusammengetragen und katalogisiert. Durch speziell entwickelte Algorithmen wurden die enzymatischen Reaktionen analysiert und die reaktiven chemischen Epitope identifiziert. Diese wurden verwendet, um mittels Substruktursuche neuartige Substrate für Prenyltransferasen vorherzusagen, den bestehenden Substratraum zu kategorisieren und ihn dem Benutzer, über eine Web-Oberfläche, zugänglich zu machen.

Der dritte Teil dieser Arbeit umfasst eine Fallstudie, in der die Regioselektivität von fünf Prenyltransferasen bezüglich unnatürlicher Prenylrest-Donatoren untersucht wurde. Eine detaillierte biochemische Untersuchung offenbarte, dass sowohl die Umsatzraten als auch die Anknüpfungspunkte der verschiedenen Dimethylallylpyrophosphat-Analoga von der Kombination aus der jeweils eingesetzten Prenyltransferase und dem Prenylrest-Donator abhingen. Um die Ursachen dieser Abhängigkeit auf der atomistischen Ebene zu verstehen, habe ich Molkulardynamik-Simulationen eingesetzt und die zeitliche Entwicklung der Trajektorien der Enzym-Substrat-Prenyl-Donator-Systeme

untersucht. Ich konnte zeigen, dass ein einfaches abstandbasiertes Model die Reaktivität der elektro- und nukleophilen Zentren eines Substrat-Prenyl-Donator-Paars und so die Regioselektivität der Reaktion erklärt. Diese Analyse stellt auch einen ersten Schritt dar, die Reaktivität der Prenytransferasen mechanistisch zu verstehen und unterstreicht den potentiellen Nutzen dieser Enzymfamilie als Biokatalysatoren für Friedel–Craft-Alkylierung.

Im letzten Teil stelle ich am Beispiel des β_2-adrenergen Rezeptors (β_2AR) den synergetischen Einsatz von ligand- und strukturbasierten chemoinformatischen Methoden vor. Die Kombination aus fingerprintbasierter molekularer Ähnlichkeit, substrukturbasierter Suche und Dockingstudien ermöglichte uns 13 Liganden für diesen guterforschten Rezeptor zu identifizieren. Diese Liganden, von denen zehn im Vergleich zur bekannten Ligaden als neu eingestuft werden können, wurden auf Grundlage von zwei in einer früheren Studie beschriebenen Liganden gefunden, die beide über ein einzigartiges Grundgerüst verfügten. In diesem Licht wird der erfolgreiche, aufeinander folgende oder auch parallele Einsatz von verschiedenen, sich konzeptionell unterscheidenden Screeningmethoden unterstrichen und der intrinsische komplementäre Charakter von ligand- und strukturbasierten Konzepten offensichtlich.

Contents

Preface

Sah früh von des Schloßhofes Runde hoch über die Dächer ins Land,
Wo rings durch die Auen im Grunde die Lahn flicht ihr silbernes Band.
Und feierlich wogte im Tale der Glocken tiefdröhnende Flut,
Verklingend im Dankeschorale! Alt-Marburg, wie bin ich dir gut!

aus ***Alt-Marburg, wie bin ich dir gut!****, Text und Musik: Otto Janson, 1927*

Keine langwierige Recherche, kein Abwägen des Für und Wider, keine Auskünfte, keine Studienberatung, keine Rundfahrten durch die Republik. Die Entscheidung für Marburg war eine spontane, eine Bauchentscheidung. *Geh' nach Marburg, dort ist es schön!* hieß es. Und das stimmt! Was die Wahl der Studienrichtung anging, so stand die Entscheidung für Chemie seit der Lektüre von *Faust* fest. Wenn man doch begreifen möchte, *was die Welt im Innersten zusammenhält*, bietet die Chemie eine einzigartige Möglichkeit, die große und die kleine Welt zu sehen. Vom Makroskopischen ins Mikroskopische und zurück. Vom Organischen ins Anorganische, über das Theoretische bis ins Physikalische.

Doch es reichte nicht eine Studienrichtung zu wählen. Während des Studium musste ich noch eine weitere Entscheidung im Hinblick auf die chemische Disziplin treffen. Denn seit Goethe ist viel Geschehen und die Naturwissenschaften im Allgemeinen, und die Chemie im Besonderen, sind zu einem interdisplinären Geflecht aus sich überschneidenden Teilbereichen geworden, die in ihrer Tiefe und Komplexität kaum überschaubar, geschweige denn in der Gänze begreifbar sind. Früher oder später, doch oft spätestens dann, wenn man sich auf die Suche nach einer geeigenten Arbeitsgruppe für die Anfertigung der Bachelorthesis macht, muss die chemische Disziplin gewählt werden. So entschied ich mich für die computergestützte Wikstoffentwicklung – die für mich rückblickend uneingeschränkt richtige Wahl.

Doch zunächst wusste ich nur, was ich nicht tun wollte: Organische und anorganische Synthese, beides prominente Betätigungsfelder Marburger Chemiker und durchaus spannend und reizvoll, jedoch hinter physikalischer und theoretischer Chemie, was mein persönliches Interesse anbelangte, zurückbleibend. Erst im Herbst 2009, kurz vor dem Ende des Hauptstudiums, kam es für mich zu einer wegweisenden Begegnung. Im Zuge der Vorlesung zur Theoretischen Chemie besuchte ich das *4th Rhein-Main Molecular Modelling Meeting*, welches von Prof. Guido Germano organisiert wurde. Dort hörte ich die Vorträge von Prof. Gerhard Klebe und Prof. Gernot Frenking, die sich zwar in ihrer Thematik unterschieden – Vorhersage von Affinität von Protein-Ligand Komplexen auf der einen und Moleküle mit ungewöhnlichen Bindungsverhältnissen auf der anderen Seite – mir jedoch auf eine spannende und faszinierende Art zeigten, dass komplexe Fragestellungen der modernen theoretischen und medizinisch-pharmazeutischen Chemie mit Hilfe des Computers erforscht werden können. Ein wahrlich horizonterweiterndes Erlebnis für einen Studenten, der nach Marburger Manier die meiste Zeit im Labor verbrachte.

Prof. Klebes Vortrag und die Tatsache, dass die Wechselwirkungen von Proteinen und kleinen Molekülen computergestützt untersucht und modelliert werden können – und das stets im Kontext des größeren Bildes der Wirkstoffentwicklung – fesselten mich sofort. Es war fortan mein Wunsch, auf dem Gebiet der computergestützten Wirkstoffentwicklung zu lernen und zu forschen. Umso erfreuter war ich, als ich meine Bachelorarbeit im Arbeitskreis von Prof. Klebe und dort unter der Betreuung von Tobias Craan anfertigen durfte. Es waren sehr lehrreiche und produktive Wochen, während dieser meine Begeisterung für Biochemie, Pharmazie und Strukturbiologie weiter stieg. Ich möchte hier Tobias für seine Geduld und hingebungsvolle Unterstützung meiner Tätigkeit in einer für mich völlig neuen Arbeitsumgebung danken. Die Bachelorarbeit war in vielerlei Hinsicht für meinen weiteren akademischen Werdegang bestimmend: Es war mir klar, dass ich der Wirkstoffentwicklung am

Computer treu bleiben würde und dass ich meine Masterarbeit ebenfalls auf diesem Gebiet anfertigen würde. Die Verflechtung der *in vitro* und *in silico* Welten zu einem sich ergänzenden und unterstützenden synergetischen Zusammenspiel stellte für mich schon damals ein attraktives Betätigungsfeld dar. Ganz im Sinne der faust'schen Freude am Osterspaziergang – *Hier bin ich Mensch, hier darf ichs sein* – erfüllte mich die Arbeit auf diesem Gebiet mit immerwachsender Neugier und Begeisterung.

Es war nur logisch, dass ich auch mein ganzes Masterstudium nach diesem Thema ausrichten wollte. So begegnete ich Personen, die meine Begeisterung für Strukturbiologie und Computerchemie weiter anfachten. Ich führte in den Arbeitsgruppen von Prof. Mohamed Marahiel und Prof. Gernot Frenking Forschungspraktika durch, die mich zusammen mit den begleitenden Vorlesungen zur Biochemie, Biokatalyse und Theoretischer Chemie auf die Arbeit mit Proteinen, kleinen Molekülen und natürlich Computern bestens vorbereiteten. Erste Gehversuche auf dem Gebiet des Programmierens sollten sich ebenfalls als vorteilhaft herausstellen. Im Spätherbst 2011, zwei Jahre nach dem ersten Kontakt mit Prof. Klebe, schrieb ich ihn nochmals an. Diesmal mit der Anfrage nach einem Platz zur Anfertigung einer Masterarbeit. Meine Enttäuschung war zunächst groß, als Prof. Klebe mir keinen freien Platz anbieten konnte und ich mangels Alternativen ziemlich nervös wurde. Doch Prof. Klebe ließ mich nicht einfach ziehen. Er verwies auf einen jungen Nachwuchswissenschaftler, der damals im Begriff war, seine Arbeitsgruppe aufzubauen, und auch Interesse an einem Masteranden hatte. Ich sollte doch kurz bei ihm vorbeischauen und mich vorstellen. Und so klopfte ich an die Tür von Dr. Peter Kolb, den Betreuer meiner Masterarbeit und späteren Doktorvater.

Zwei Doktoranden, ein Post-Doc und zwei Büroräume. Das waren die Kennzahlen der Arbeitsgruppe Kolb, als ich im Frühjahr 2012 zur Anfertigung meiner Masterarbeit dazustieß. Vom ersten Tag an war Peter mehr als nur ein Betreuer, der meine wissenschaftliche Arbeit beaufsichtigte. Er gab mir alle Freiheiten, meine Projekte während der Masterarbeit und später als Doktorand selbstständig zu entwickeln, eigene Ideen einzubringen und sie auch kritisch zu hinterfragen. Von unschätzbarem Wert waren auch seine eigenen Ideen und Hilfestellungen, gerade dann, als ein Projekt in einer vermeindlichen Sackgasse steckte. Es ist Peters Gabe, das größere Bild zu sehen. Dieser Blick über den Teller- oder (besser) Bildschirmrand half mir oft, aus einer verfahrenen Situation herauszukommen. Seine Weitsicht hat mich neben seiner Geduld, großen Kreativität, steten Aufmerksamkeit und persönlichem Engagement tief beeindruckt und nachhaltig geprägt, fachlich wie auch persönlich. Ich kann mich nur für alle seine zukünftigen Doktoranden freuen. Ich bin mir sicher, dass derer noch viele folgen werden. Vielen Dank Peter.

Dank gebührt selbstverständlich auch meinem Zweitgutachter Prof. Shu-Ming Li, ohne den meine Arbeit nicht hätte entstehen können. Besonderen Dank möchte ich ihm für die allzeit angenehme und effektive Zusammenarbeit aussprechen, die Grundlage des Gelingens meiner Forschungsprojekte war. Nicht zuletzt sein Enthusiasmus für computergestützte Methoden und unsere in zahlreichen Diskussionen entwickelten Gedanken befähigten mich, meine Ideen in praktische Anwendungen umzusetzen. Im Zuge dessen danke ich auch allen meinen Kollaborationspartnern, die mit mir auch über meine Promotion hinaus an mannigfaltigen Projekten beteiligt sind. Der durchweg fruchtbare Austausch von Ideen und Anregungen, die unvoreingenommene Sicht auf die Wirkstoffentwicklung mit Hilfe des Computers sowie das reibungslose Zusammenspiel von computergestützter Vorhersage und dem experimentellen Beweis war und ist maßgeblich für den Erfolg meiner wissenschaftlichen Arbeit. Im Einzelnen möchte ich meinen Dank richten an Jillian Baker, Jean-Louis Reymond, Daniel Rosenbaum, Katja Backhaus und Peter Gmeiner. Es bereitete mir eine große Freude, interdisziplinär und international Wissenschaft zu betreiben, mich mit ihnen auf Konferenzen zu treffen, auszutauschen und zu vernetzen und Projekte gemeinsam voranzubringen.

Ich möchte an dieser Stelle meinen herzlichen Dank an Prof. Klebe richten, der mein Interesse für die Chemoinformatik weckte und mir den Start in das wissenschaftliche Arbeiten ermöglichte. Auch später, im Zuge meiner Master- und Doktorarbeit, unterstützte er mich bei meinem Werdegang als Wissenschaflter, sei es durch den stets geförderten wissenschaftlichen Austausch oder ehrliche und konstruktive Kritik. Nicht zuletzt möchte ich Prof. Klebe für das mir entgegengebrachte Vertrauen danken, auf dessen Grundlage er mir die Bekanntschaft mit Peter ermöglichte. An gleicher Stelle möchte ich herzlich Lydia Hartleben danken für ihre Hilfe bei allen administrativen und organisatorischen Fragestellungen, die mir während meiner Tätigkeit am Fachbereich begegneten. Ich danke ihr für ihre Geduld und Hilfsbereitschaft, die sie mir und meinen Fragen, Formularen, Abrechnungen und Anträgen stets entgegenbrachte.

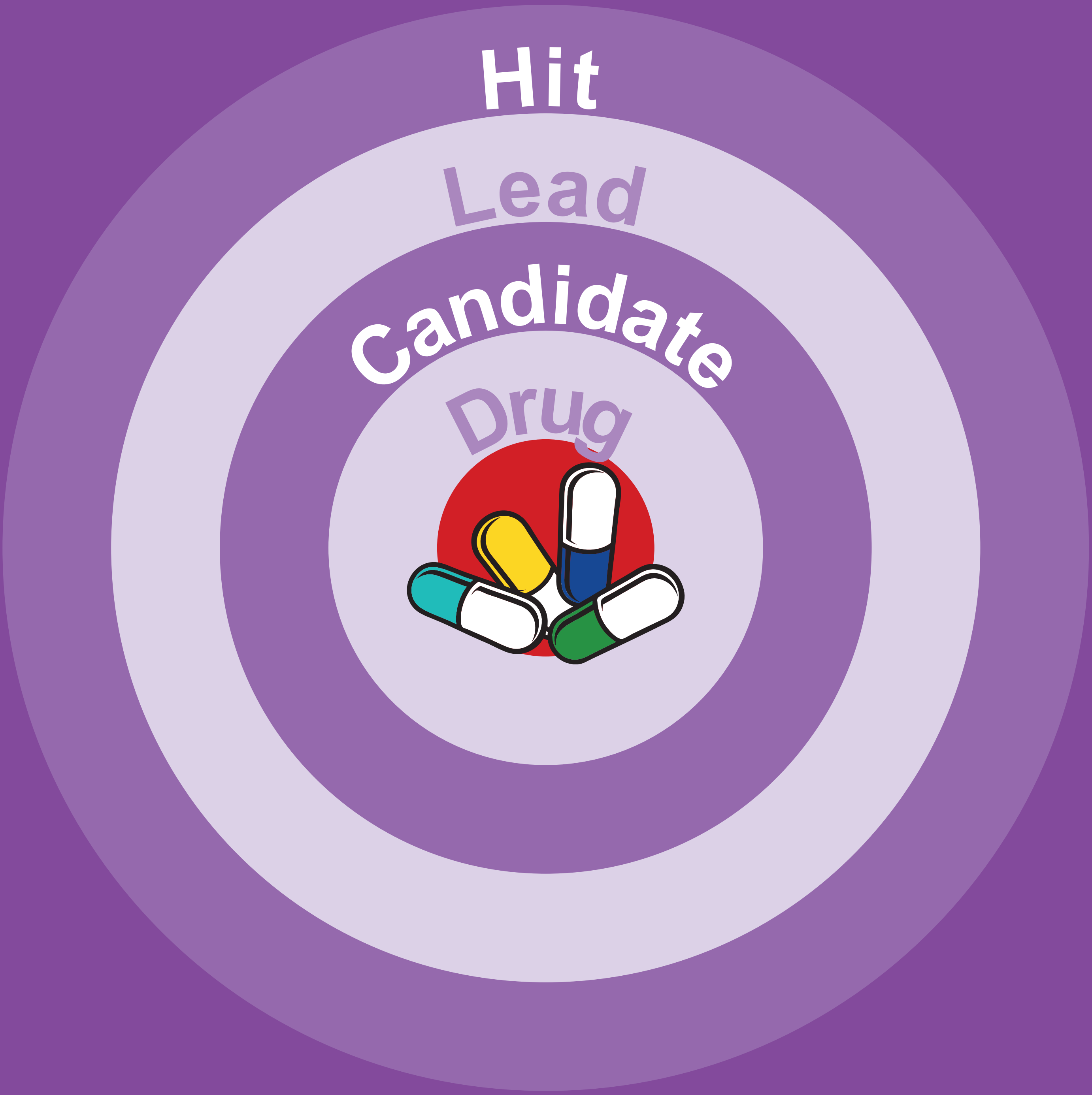
Arbeit in einer Forschungsgruppe kann nur von Erfolg gekrönt sein, sofern die Arbeitsatmosphäre und der kollegiale Zusammenhalt stimmen. In dem Zusammenhang kann ich mich nur glücklich schätzen, dass ich mit so zuvorkommenden, freundlichen, hilfsbereiten und gleichermaßen professionellen Kollegen zusammen arbeiten durfte. Ich danke Maria Martí-Solano, Anja Flöser, Magdalena Scharf, Denis Schmidt, Florent Chevillard, Corey Taylor, Matthäus Drabek, Peter Wengenmayer und Steffen Hartmann für die großartige und spannende Zeit in der Arbeitsgruppe Kolb. Sie haben direkt oder indirekt dazu beigetragen, ein produktives, erkenntnisförderndes Klima zu schaffen, in dem der wissenschaftliche Austausch eine Freude war. Im gleichen Zuge möchte ich meinen Dank an Prof. Steinmetzer und seine Arbeitsgruppe richten, im Einzelnen Jenny Kouretova, Aline Keils, Kornelia Hardes, Anton Epp und Benjamin Wenzel. Zur wissenschaftlichen Arbeit in der Akademia gehört auch die Lehre und es war eine große Freude für mich, gemeinsam mit der Arbeitsgruppe Steinmetzer daran zu arbeiten, die praktische und theoretische Ausbildung von Studienanfängern auszugestalten und langfristig zu verbessern.

Nicht selten beginnt ein neuer Lebensabschnitt, bevor der vorige abgeschlossen werden kann. So überlappte auch meine Promotionszeit meinen Einstieg ins Berufsleben. Als das dritte Promotionsjahr fortgeschritten war und das Ende der Promotion sich vage in einer nahenden Zukunft abzeichnete, hatte ich den Wechsel in ein pharmazeutisches Unternehmen vollzogen. Dies war gewollt und willkommen, jedoch gleichwohl ein radikaler Schritt. Gut die Hälfte der in dieser Arbeit vorgestellten Studien befand sich zu diesem Zeitpunkt erst in einem frühen Stadium der Verschriftlichung. Das Abreißen der Zelte in Marburg, neuer Wohnort, neuer Arbeitsplatz, neue Aufgaben und noch nicht einmal fertig promoviert – das bedeutete Stress und davon nicht wenig. Ich konnte und kann mich daher glücklich schätzen, an meinem neuen und jetzigen Arbeitsplatz mit Kollegen zusammenzuarbeiten, die mich vom ersten Tag an voll unterstützten. Mein Dank geht an Mireille Krier, Michael Krug, Christian Herhaus, Paul Czodrowski, Daniel Kuhn, Tim Knehans, Fabian Steinmetz und meinen Vorgesetzten Friedrich Rippmann für ihre Unterstützung während des Wechsels von der Universität in die Pharmaindustrie. Sie halfen mir tatkräftig dabei, beide Aufgaben- und Verantwortungsfelder unter einen Hut zu bringen. Ihr moralischer Beistand, ihre Zuversicht und das Vertrauen in meine Fähigkeiten waren gerade in der letzten Phase der Anfertigung dieser Arbeit von besonderem Wert. Vielen Dank CCB-Team.

Dank gebührt zu guter Letzt meiner Familie und meinen Freunden, die immer an meiner Seite waren, immer ein offenes Ohr für mich hatten und zu meiner Arbeit ein wichtiges Gegengewicht darstellten. Durch ihre unermüdliche Unterstützung und Ermutigung konnte ich in schwierigen Phasen neue Kraft schöpfen. Die Zeit, die ich mit ihnen im Zuge meines Studiums und sodann meiner Promotion in Marburg verbringen durfte, möchte ich um keinen Preis missen – sie wird mir immer als eine besonders schöne Erinnerung im Gedächtnis bleiben.

Und bin ich ein alter Geselle und bleichte die Zeit mir das Haar,
So such ich noch einmal die Stelle, wo damals so glücklich ich war.
Seh still auf das alte Nest nieder und schwing dann jungselig den Hut
Stimm ein in das Lied aller Lieder: Alt-Marburg, wie bin ich dir gut!

*aus **Alt-Marburg, wie bin ich dir gut!**, Text und Musik: Otto Janson, 1927*


Hit
Lead
Candidate
Drug

Introduction

This chapter of my thesis is meant to outline the drug discovery and design process that can be found in a more or less conserved form throughout the pharmaceutical industry. I decided to exploit this long, expensive, multi-phase process as a stage for the presentation of the chemoinformatic methodology that I heavily used in all of the projects described in this thesis.

In this introduction I describe selected methods, their conceptual design and how they impact the early drug discovery and design stages. It is my intention to not only emphasize strengths of tools such as molecular docking and molecular similarity but also to allude to their weaknesses and, eventually, to lead over to alternative approaches I developed and which form a major part of this thesis.

Computer aided drug discovery and design

There are many aspects of drug discovery that can be addressed to increase its lower than expected productivity.

György M. Keserű and David C. Swinney, A personal Foreword to ***Thermodynamics and Kinetics of Drug Binding****, 2015 Wiley-VCH*

In this short statement Keserű and Swinney laid out the real nature of the drug discovery and design process: It can be seen – from its early beginning to its very end – as a multidimensional optimization problem. A problem whose many aspects, facets and questions – dimensions – can rarely be dealt with, understood or answered one-by-one. On the contrary, the inter-dependency of the properties of a chemical entity such as a molecule or an antibody, in the context of its desired action in the human body, compels the researchers to simultaneously work on a variety of properties. Not uncommonly, these properties are inter-connected in such a way that one feature can only be improved by impairment of a different one. Thus, the overall success rate of a drug discovery and design campaign depends strongly on the starting chemical entity, its initial set of properties and whether they can be further adjusted to yield a desired state.

This choice of the initial entity, a promising molecule, a needle in a haystack, is crucial. Even more so, considering that the process of bringing a new drug from discovery to market takes many years (12-14 years), up to \$1.2-\$2.6 billion dollars (1,2) and fails with a probability of over 90 %. (2) The high chance of failure, the ever-incrementing costs during the campaign, peaking at the very last stages of clinical development (3), necessitates robust and reliable methodology for the identification of the aforementioned needle already at the beginning of the long process. In this early stage of drug discovery and design, but not exclusively there, *in silico* methods can truly shine and deliver on the identification and optimization of promising starting points. (4,5,6,7) Among many methods and concepts in computer-aided drug discovery and design (CADDD), molecular similarity and molecular docking are widely and successfully applied to the search for the needle within a vast chemical space. In addition, developments with respect to chemical reactions, predictions of reactivity and applicability of chemoenzymatic reactions open new avenues to optimization of hit matter – novel approaches to a yet-to-be uncovered chemical space.

1. CADDD along the drug discovery and design pipeline

The long and costly process of bringing a new drug to the market consists of a multitude of sequential discovery and design phases, decision points and milestones (**Figure 1**). The general layout of the drug discovery and design pipeline is conserved throughout the pharmaceutical industry. Within the pipeline, there are many opportunities for bio- and chemoinformatic methodology to constructively contribute to the overall success of the campaign. (8,9,10) Indeed, the process of drug discovery has undergone revolutionary changes since the advent of technologies such as genomics, proteomics, bioinformatics, (11) combinatorial chemistry, virtual screening and *in silico* prediction of absorption, distribution, metabolism, excretion and toxicity end-points (ADMET). (12)

1.1. The concept phase

In its earliest stage – the concept phase – the molecular target has to be identified. The target, in a simple case a single protein, often an enzyme, has a key function within a metabolic or signaling pathway. The inhibition of this key function by a small chemical entity blocks the corresponding pathway and propagates all the way to the desired phenotypic effect. An example is the inhibition of the DNA damage repair machinery in tumor cells which eventually sensitizes them to radio- and chemo-therapy. Target identification mainly falls

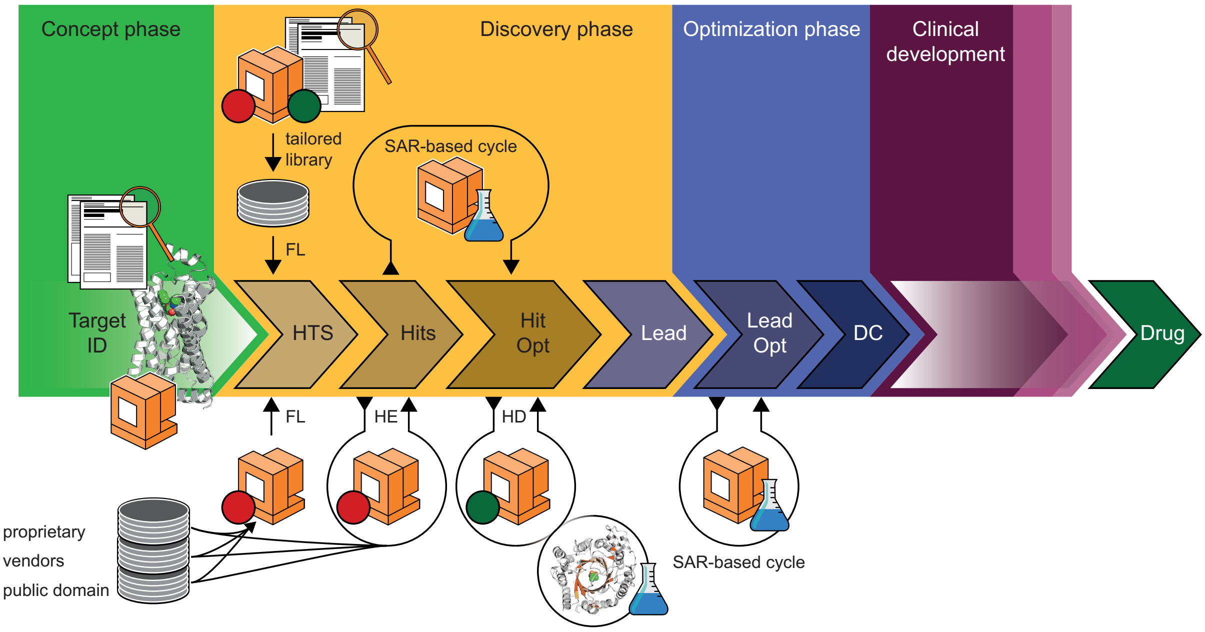


Figure 1: CADDD in perspective with the early drug discovery and design pipeline. Bioinformatics-supported target identification (Target ID) and characterization (green). Literature-based tailored library design and virtual screening prior to HTS, front-loading (FL). Hit-matter-based secondary virtual screening – hit expansion (HE) via analog-by-catalog approach. Hit diversification (HD) via novel chemistry routes, *e.g.*, chemoenzymatic transformation leading to optimized hit-matter (Hit Opt) (orange). Understanding of the SAR allows for rationale design. Lead optimization (Lead Opt) and development candidate entering the clinical development phase (blue). Red circles: entry points for molecular similarity approaches, *e.g.*, RedFrag. Green circles: Chemical space exploration with reactivity prediction tools, *e.g.*, PrenDB and SAR-by-Enzyme.

within the domain of Biology, where a particular pathway is characterized. Mining of available biomedical data driven by bioinformatics has led to a significant increase in identified targets. Patents, gene expression data, proteomics, examining of mRNA synthesis and protein expression levels and phenotypic screens contribute to the large data pool from which novel targets can be elucidated. (13)

The linking of a target – and its inhibition by a small molecule – to a therapeutic effect in a disease state is a necessary condition for target identification. Thus, target validation, the usage of orthogonal experimental approaches to prove the target's therapeutic significance and its cellular function prior to a screening campaign, is a crucial requirement. Still, of similar importance is the target's druggability: its potential to be modulated by a yet-to-be-discovered drug molecule that upon binding triggers the desired biological response *in vitro* and eventually *in vivo*. In this respect chemoinformatics contributes early on in the process. Based on available structural data, if necessary based on homology models in case no X-ray data is available, putative binding pockets can be analyzed. Based on the curvature of the proteins' surface, its depth and enclosure, relative number of polar, apolar and ionizable amino acids within detected cavities and number and nature of putative hotspots contribute to the estimation of the propensity of ligand binding. (14) A variety of algorithms has been developed for that purpose, *e.g.*, DoGSite (15), FTMap (16), SiteMap (17) and MOE SiteFinder (18).

1.2. The discovery phase

Once a target is identified and validated, *i.e.* its significance for a pathway and therefore an indication is proven, chemoinformatic methodology plays a distinct role prior to high-throughput screening (HTS): The main contribution is the feasibility assessment and the design of targeted libraries or of large-scale pre-screens of molecular databases. These early screen and design steps can be accomplished by a virtual high-throughput screening (vHTS). Here, based on available information about the target's structure, its known ligand space, or data derived from analogous systems, entities from multi-million-sized databases of chemical matter can be probed for their likelihood of interacting with the underlying target.

Virtual screening techniques are not only applicable prior to the large-scale HTS. Given a set of molecules – dozens or even hundreds of hits from the HTS – a second round of vHTS can be conducted. This analog-by-catalog approach, where experimentally verified hits that features molecular characteristics beneficial for bioactivity are exploited and used as query for screening, is a fast and efficient method for the discovery of low-hanging fruits and an opportunity for early hit expansion.

As the number of identified hits and the amount of experimental data grows, chemoinformatic methodology proves to be helpful in order to analyze structure-activity relationships (SAR) of the hits and build binding models and activity hypotheses. These models are essential for a guided, rational design of novel molecules. Their testing *in vitro*, the reevaluation of the experimental data, its incorporation into new models, and design of further molecules resembles an iterative hit optimization and hit expansion cycle. Eventually, a hit molecule, often containing a characteristic chemical scaffold, emerges from the hit optimization cycle: Its favorable properties, *e.g.*, high affinity and selectivity towards the target, award it lead status.

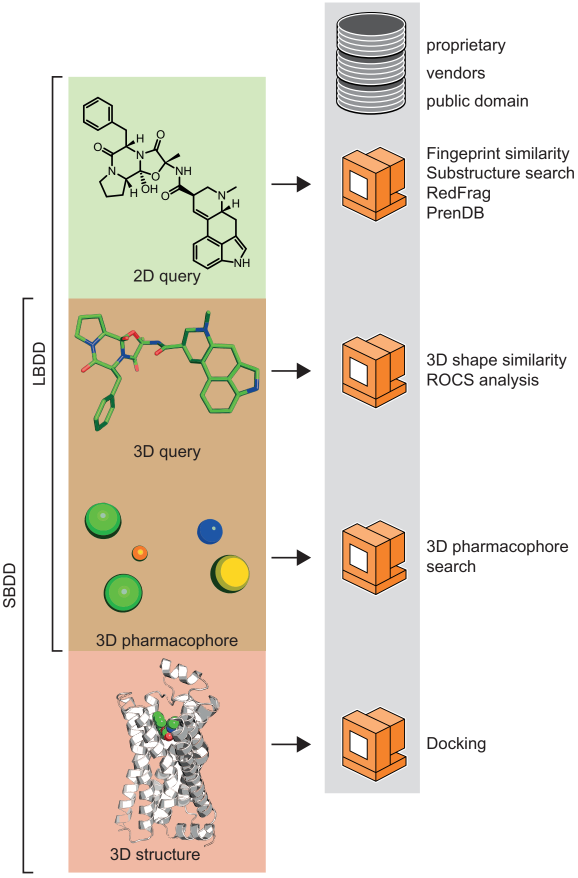


Figure 2: According to available data, CADDD methodology can be used to supplement a project with novel chemical matter: Given only a reference compound, 2D molecular similarity in its different flavors can be used. With gathering of structural information, the spectrum of available methods grows. Three-dimensional shape comparison and structure-derived pharmacophore searches (hybrid methods of LB- and SBDD) and eventually docking becomes possible.

1.3. The optimization phase

A lead compound is a chemical entity which evolved during the discovery phase from among several hundreds of hits. It prevailed because of its unique set of properties: affinity towards the target in the nanomolar range, a balanced solubility and permeability profile, promising efflux ratio and low (calculated or modeled) off-target affinity, to name a few. In the lead optimization process properties like metabolic stability, pharmacodynamic and pharmacokinetic profile, potency and selectivity are further and thoroughly improved. The goal of the optimization phase is a compound or compound family that fulfills the requirements of being tested *in vivo* and becoming a development candidate (DC).

2. Ligand- and structure-based drug design

The multi-step drug discovery and design pipeline – target identification and validation, hit discovery and optimization, lead identification and optimization, candidate selection – offers a variety of entry points and opportunities for chemoinformatic *in silico* methodologies.

Already at the early stage, shortly after target identification and characterization and before the HTS, ligand- and structure-based approaches play an important part in increasing the chance to discover promising hits. Depending on the available information, structural data of the target, size of its ligand space, existence of analogue systems, a multitude of orthogonal methods and – beneficially – its combination can be applied (**Figure 2**). Molecular similarity methods and molecular docking are fast and robust methods that can supplement the discovered hits with additional compounds, both of which require little information about the system of interest. Similarly, the field of quantitative structure-activity relationship (QSAR) and its predictive tools help to predict physico-chemical endpoints based on 2D and 3D structural information of the so-far-explored chemical space. Pharmacophore models – knowledge-based binding hypothesis filters – incorporate the three-dimensional arrangement of chemical epitopes crucial for binding, allow for fast compound classification but also requires elaborated data.

All of them have in common that they can be used to screen proprietary, commercially available and public domain databases of small molecules, thus reducing the available chemical space (several hundred thousands to many millions compounds in size) to a fraction of mere thousands of putative binders or virtual hits. Popular resources for screening compounds are, among others, the ZINC database curated selection of commercially available compounds (19); the National Cancer Institute Library (NCI); the ChEMBL database of small molecules and their reported binding functional effects to macromolecular targets (20); and the PubChem database maintained by the National Center for Biotechnology Information (NCBI) (21).

The usage of ligand- and structure-based methodology in drug discovery and design is not limited to large-scale virtual screenings (VS). Within the discovery and design pipeline, there are many interfaces where molecular similarity and/or docking can be beneficially applied. Most importantly they contribute to the understanding of the SAR. Once a ligand-receptor complex structure has been solved and a sufficient number of closely related analogs has been synthesized and their affinity towards the target determined, molecular modelling can be used to correlate the observed affinity values with structural information. Even without an existing ligand-receptor complex, molecular docking can be used to generate a binding mode hypothesis that serves as seed for the posing of a ligand series into the receptor and the analysis of their interactions.

2.1. Molecular similarity

Within the ligand-based method family, molecular similarity and its vast – and still growing – number of different incarnations is a very popular screening technique. (22) Indeed, it is one of the most heavily explored and exploited concepts in chemical informatics. (23,24,25)

Molecular similarity and the underlying similarity property principle (SPP), which states that similar compounds should have similar properties and most of all similar biological activity (26), seems intuitive and simple, but only at the first glance. **Figure 3** shows four geometric shapes which, to the human eye and to different extents, share similar properties: Obviously, the hexagon and the square share identical surface and outline color. The same is true for the circle and the triangle. Concerning the number of angles, the triangle is more similar to the square than the hexagon and the circle is least similar to the other three figures. On the contrary, the relative high number of angles in the hexagon resembles the shape of the circle more accurately. Finally, all shapes are identical with respect to their surface area.

This example illustrates the complexity of the question how similar objects are and based on which properties and features of a given set of objects similarity can be argued about. These questions get more and more intricate the more abstract and complex an object set becomes. Potentially bioactive molecules, a set as large as 10^{60} objects (27,28), assembled from a dozen of chemical elements and connected by a handful of bond types, pose delicate case where concepts of similarity are as indispensable as they are multifarious.

They are indispensable because otherwise navigation through the chemical space that we have explored so far, both synthetically and virtually, would be limited to the closest analogs

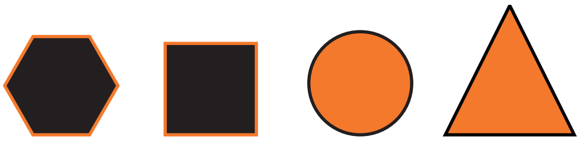


Figure 3: Similarity depends on the properties that serve as basis for comparison. Color, shape, surface area, number of angles are only a few, based on which the similarity of the four geometric shapes can be assessed.

for a given query compound. Small, systematic modifications of the molecular skeleton, which lie in the reach of medicinal chemists, would predominate the chemical space built around a given hit or parent structure. Larger leaps into – measured by human intuition – more distant chemical environments are only possible if an automated and fast method for similarity assessment is invoked. Such a method compares a molecule with millions of other chemical entities calculates the similarity between them and fetches the most similar – or dissimilar – according to pre-defined criteria.

The simple example with squares and triangles showed that similarity assessment is not straightforward and even less so if computers, intuition-free devices, are entrusted with the task. Similarity calculation has many flavors and the pool of molecular similarity methods is steadily growing, as is natural for a concept that deals with objects as complex as molecules. Furthermore, researches constantly modify and extend present methods and add new concepts to the field. Given also the fact that a computer knows nothing about similarity but is very efficient in evaluating formulae, the manifold concepts of similarity have to be translated into mathematical fabric, which in turn is many-faceted and thus prone to creative interference.

Manifestations of molecular similarity are diverse. Conceptually, one could differentiate between one, two- and three-dimensional methods. One-dimensional similarity or chemical similarity relies primarily on physico-chemical characteristics of molecules, *e.g.*, solubility, lipophilicity, molecular weight, etc. Properties that can be derived from the chemical formula but in some cases also necessitates the molecular structure. In many cases these properties themselves are subject to calculations based on the molecular structure: LogP values, dipole moments, kinetic and thermodynamic solubility and further pharmacokinetic and pharmacodynamic end points to name a few. (22) Two-dimensional similarity relies on information derived from the molecular graph (29), *i.e.*, the plain structure of the molecule, *e.g.*, shared substructures, composition of ring systems, distribution of torsional angles, topological arrangement of chemical epitopes, etc. A direct comparison of molecular graphs or parts of it – graph and subgraph isomorphism – are computationally demanding and not as widely spread as other implementations of 2D similarity. Still, substructure searches, where a prominent part of the molecular graph – often the scaffold – is used as query and is searched within a large number of target molecular graphs, is very popular, especially in the analog-by-catalog hit expansion approach. More widely applied are fingerprint-based methods that use precalculated vector representations of molecules such as extended-connectivity or circular fingerprints (ECFP (30), FCFP (31)) In a fingerprint, information about chemical constitution and structural features (32) of a molecule can be stored (mostly in a binary fashion of *on* and *off* bits). (32,30) Absence or presence of distinct structural features, rings, functional groups, H-bond donor and acceptor functionalities, are encoded in the fingerprint, effectively abstracting the molecular graph into a one-dimensional bit-string. Fingerprints, conceptually, are used to calculate the global similarity of a pair of molecules, *i.e.*, by means of the SPP, a compound that resembles a potent query molecule structurally and physico-chemically should show similar bioactivity. Additionally, and in contrast to pharmacophore-based local similarity approaches, fingerprint-derived similarity does not rely on specific knowledge about structural features important for activity. Thus, they are applicable where only little is known about the target and its ligand space. However, although fingerprint-based methods are efficient in screening large databases, they lack the ability to efficiently retrieve compounds with different scaffolds – key or core structural motifs. (33,34) A similarity calculation based on such bit-strings along with their creation is handled efficiently by computers. Together with a variety of similarity coefficients – mathematical formulae that evaluates the numerical similarity based on two input fingerprints – large chemical spaces can be screened and explored.

Three-dimensional similarity (35) in comparison with 2D similarity and its fingerprint-based abstraction seems, at the first glance, more intuitive: Molecules are three-dimensional arrangements of atoms, thus to discard information about their 3D shape has to be disadvantageous in light of the SPP. Indeed, comparison of molecules in 3D can lead to more accurate similarity assessment, even more so through the fact that two compounds with high 2D similarity do not necessarily occupy similar conformational states in a biologically relevant scenario. Still, 3D similarity assessment has at least three major drawbacks that renders its application limited: i) In order to obtain reasonable results, the 3D conformation of the query molecule in its biologically active state has to be known. This is only rarely the case, especially in the early stages of the drug development and discovery. ii) The necessary overlay of structures in three dimensions is computationally far more demanding than in 2D and necessitates the enumeration of conformations of the target molecules. iii) Chemists are trained on the basis of plain molecular graphs and in general more comfortable with 2D-driven analyses. (22)

Nevertheless, three-dimensional similarity assessment methods – pharmacophores, ROCS analysis, etc. – bear the desired potential of scaffold hopping. (36) A scaffold hop is considered to have occurred when a molecule differs in its characteristic core chemical epitope, scaffold or substructure, from the query or parent molecule but still resembles the spatial arrangement of decorative (crucial in terms of function) substituents. Scaffold hops are of great importance in pharmaceutical industry in terms of intellectual property (IP) and the related strategy of patent breakage. From this point of view 3D similarity methods seems attractive and – *vice versa* – scaffold hopping-capable 2D-based similarity concepts. **Part 1**, a chimeric similarity approach that combines fingerprint-based and graph isomorphism-driven methodology, describes such a concept called **RedFrag**. There, the holistic view of compounds is abondoned as they are abstracted to their reduced, fragment-featured local representation and compared thereafter. In contrast to other feature-, fragment- or structural motif-based molecule abstraction, RedFrag does not require a large pre-defined dictionary of, *e.g.*, SMARTS-encoded substructures or features. It relies solely on the definition of only a few bonds that separate structural motifs within a molecule.

2.2. Molecular docking

Methodology that makes use of three-dimensional structural information gathered from biological targets – via X-ray crystallography or by nuclear magnetic resonance (NMR) – is a prominent and widely applied component of hit discovery and design throughout the early and middle stages of the discovery and design pipeline. Structure-based design – and there molecular docking specifically – benefits from the large and steadily growing number (more than 100,000) of resolved structures of macromolecular drug targets. (37) Molecular docking, its large-scale application in virtual screening campaigns, and molecular dynamics (to a smaller extent due to its computationally demanding nature) are popular and frequently applied. (38)

Since its advent in the 1980s, molecular docking became an essential tool in drug discovery. (39) Its ability to predict the conformation of a small-molecule ligand within the binding site of the receptor renders it uniquely useful for understanding ligand-receptor interactions. In order to screen the conformational space of the ligand within the context of the binding site of the receptor, docking tools employ two strategies: i) In a systematic manner the structural parameters of a compound, *e.g.*, torsional, rotational, translational degrees of freedom, are incrementally varied and the resulting conformation is scored at each step. (40) ii) Rather than systematically, the structural parameters can be modified randomly or stochastically. (41) In this case, the algorithm generates conformational ensembles and thus covers a wide range of the conformational space. In comparison to the systematic approach, stochastic sampling is more likely to find a conformation corresponding to a global energy minimum. On the contrary, the systematic approach, due to the limited coverage of the energy landscape, can be trapped in a local energy minimum. Still, the more extensive sampling of the former comes with a higher computational demand. Exemplarily for systematic conformational sampling are tools such as FRED and HYBRID (42), DOCK (43), GLIDE (44) and FlexX (45). Stochastic searches are conducted, among others, by GOLD (46), MOE (47) and AutoDock (48).

In addition, molecular docking algorithms also execute quantitative estimations of binding energetics commonly known as scoring. (39,49) This score – generally a probability that a created pose is likely to resemble the true binding geometry – allows for ranking of docked compounds according to their likelihood of binding to the target. The corresponding scoring functions are mathematical formulae that evaluate the most important physico-chemical phenomena involved in ligand-target binding, *e.g.*, intermolecular interactions (polar, charge-assisted, dispersion interactions, etc.), protein and ligand desolvation and entropic effects. (50)

Docking, viewed from the present and considering the steadily growing computational resources at our disposal, became one of the *fast* and *cheap* virtual screening methods. Its usage is only limited by the availability of structural data. However, the lack of structural information about the target can – to some degree – be compensated by homology modeling. It is therefore not surprising that docking supplemented the studies reported within this thesis: In **Part 2**, docking featured in the last step of a hierarchical screening cascade, reducing the pool of putative prenyltransferase substrates to those compatible with the active sites of the selected enzymes. In a related study on the regioselectivity of prenyltransferases, docking-generated poses were used as seeds, starting geometries, for molecular dynamics simulations (**Part 3**). Lastly, in **Part 4**, a combined fingerprint similarity, substructure search and docking study helped us to find novel ligands for the β2-adrenergic receptor.

3. Diversification and expansion of hits

In the early stages of the drug discovery and design pipeline – shortly after the HTS – it is one of the major tasks of the joint team of medicinal chemists and chemoinformaticians to increase the number of hits. Firstly, the understanding of the SAR is crucial for a guided, rationale design of new compounds. Chemical entities with small modifications that are used as probes for the exploration of the binding site of the target are designed. Their design is based on both intuition of medicinal chemists and, more methodologically, is driven by ligand- and structure-based chemoinformatic tools. The aforementioned molecular similarity and molecular docking are widely applied in an analog-by-catalog or *de novo* fashion at this inflationary expansion stage. Secondly, by small alteration of the scaffolds of hits and iteration of their decorations, a continuous effort is made to modify the compounds' properties into their desired state.

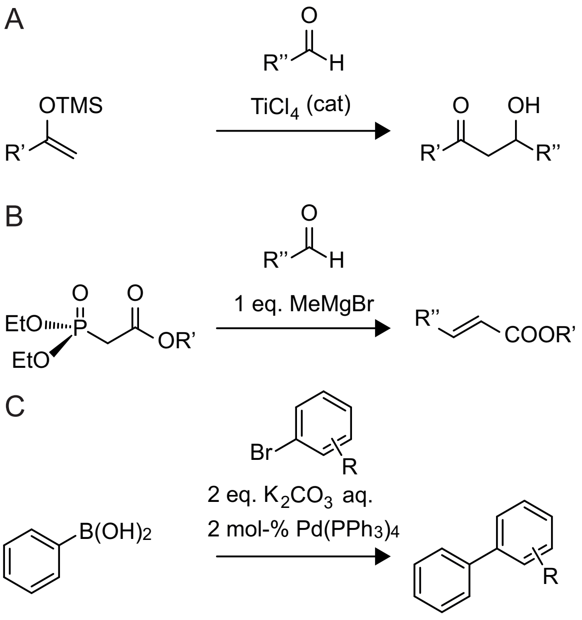


Figure 4: Three exemplary C-C bond formation reactions. **A** Mukuiyama aldol addition. **B** Horner-Wadsworth-Emmons reaction. **C** Suzuki-Miyaura coupling reaction.

In this seamlessly inter-connected phase of hit expansion, diversification and optimization the limiting factor on what can be synthetically achieved and/or easily retrieved from, *e.g.*, vendor databases, depends on the chemical space spanned by the variety of known chemical reactions: Within this chemical reaction space, arylation and alkylation of heteroatoms, acylation in general, C-C bond formation (**Figure 4**) and reduction chemistry are predominantly used in the pharmaceutical industry (they cover almost 75% of all analyzed reactions). Among the C-C bond formation reactions (11 %), cross-coupling reactions (62 %) and there the Suzuki-Miyaura

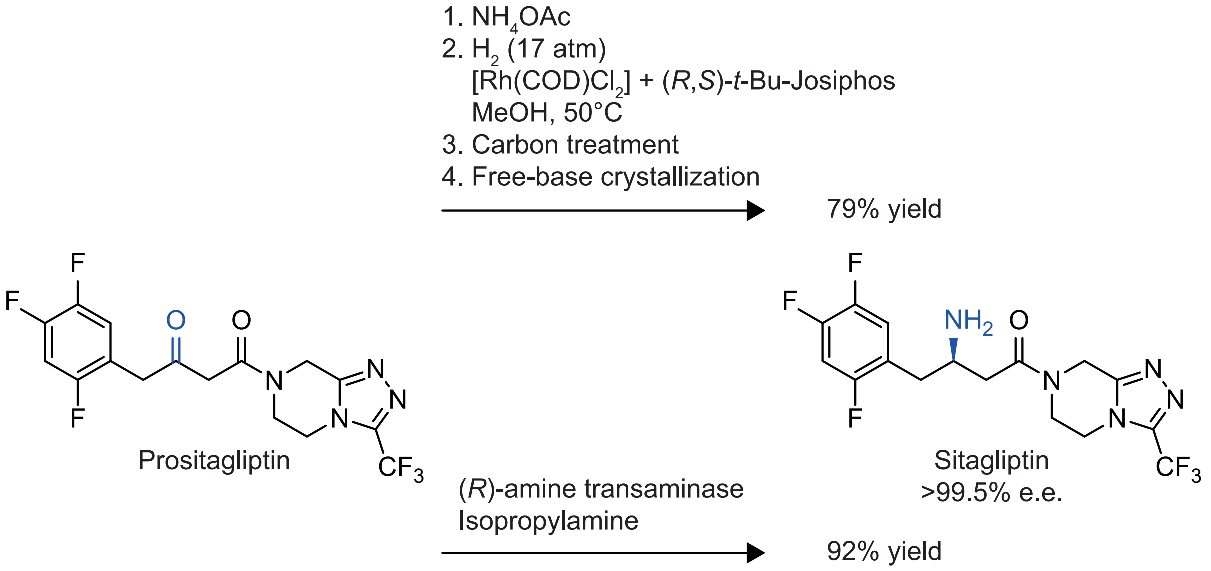


Figure 5: The enzymatic route for the synthesis of sitagliptin using an engineered amine transaminase.

reaction (40 %) takes a preponderant position. (51)

Various studies and thorough analyses described the content of the chemical space spanned by this chemical reaction repertoire: It has been shown that the chemical matter synthesized (ChEMBL database) and vast databases of computer-enumerated molecules (GDB) (52) store linear concatenations of building blocks resulting in a rod-like shaped chemical space with limited three-dimensionality. (53) It seems therefore important – even necessary – to expand our toolkit of chemical reactions, thus allowing for alternative chemistry, hit diversification in the direction of three-dimensionality, higher structural diversity and richness of molecular properties. Enzymes and their unique ability of conduct complex chemical transformations under mild conditions open a promising avenue in that direction. Efforts in understanding biocatalysts were undertaken in a study described in **Part 3** of this thesis. Joint application of docking and molecular dynamics helped us to elucidate how regioselectivity in reactions catalyzed by prenyltransferases is controlled. In order to use prenyltransferases as biocatalysts and diversifiers of chemical matter, we performed a thorough literature search and compiled the extracted data into a database of prenyltransferase reactions supplemented with predictive algorithms (**Part 2**).

3.1. Chemoenzymatic reactions – Biocatalysis

Biocatalysis is achieved by enzymes with the ability to bind and modify small molecules with regard to a determined chemical reaction. Enzymes, in contrast to other catalysts, act upon their substrates from within a mostly water-excluded cavity, the active site, that allows for unique reaction environment: Shifts in both pK_A values of the amino acids sidechains and the redox potentials of the surrounding enables enzyme to conduct complex chemical transformation under ambient conditions. (54,55) Moreover, they are asymmetric by design, thus allowing for high stereo-, chemo- and regioselectivity. (54) Biocatalysis is well aligned with principles of green chemistry: Atom- and energy efficiency, less hazardous chemical syntheses, renewable material source and prevention of waste are only few buzzwords that apply to enzymes and their use in synthetic chemistry. (56)

Although biocatalysis shines with promising features, successful applications and case studies throughout the literature, there are limitations and weaknesses. Coherently reported are three: i) The development of biocatalysts is prone to chance, slow and does not follow a pre-defined set of rules. ii) Enzymes have a limited range of stability with respect to temperature, solvents, pH value, ionic strength to name a few. iii) The number of characterized biocatalysts is still low compared to chemical catalysts. (54)

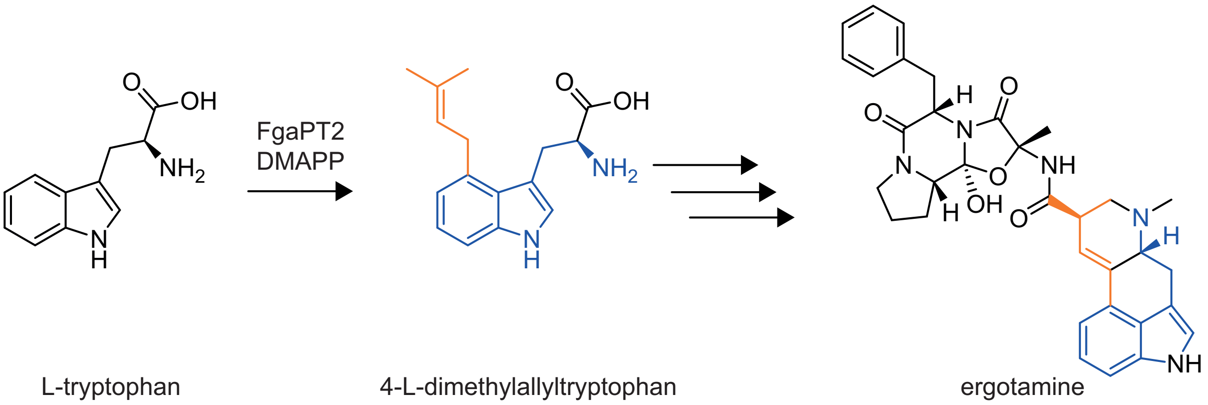


Figure 6: First step of the biosynthesis of ergotamine is conducted by the prenyltransferase FgaPT2. It attaches dimethylallyl pyrophosphate (DMAPP) to L-tryptophan. The product, 4-L-dimethylallyltryptophan, is further enzymatically modified to lysergic acid and eventually to ergotamine. Colored in orange and blue are the skeletons of the substrates of FgaPT2. Adapted from ref. (76).

In order to overcome these caveats a tremendous progress has been made throughout the decades since the dawn of recombinant DNA technology: Rational (or data-driven) protein design and combinatorial protein engineering synergistically contributed to the development of enzymes with desired properties. (57,58,59,60) Catalytic efficiency and thermal stability are of utmost importance for large-scale applications in chemical pharmaceutical industry and thus in focus of protein engineering research. In addition, enzymes with novel reaction profiles were discovered, some of them without a complement in chemistry, closing the gap to classical catalysis. For example, enzymes that catalyze named reactions such as Diels-Alder reaction (Diels-Alderase) (61), Morita-Baylis-Hillman reaction (Michael additase) (62) and Kemp elimination (Kemp eliminase) (63) has been developed.

Figure 5 shows one of many success stories of applied biocatalysis in pharmaceutical industry: The synthesis of the anti-diabetic compound sitagliptin necessitates in its last stage a rare heavy metal (Rh) catalyst in a multi-step reaction with an overall high level of waste and under harsh reaction conditions. The enzymatic route, on the contrary, applies an engineered amine transaminase, simplifies the overall reaction, increases the yield by more than 10 % and the productivity (kg/L per day) by 53 %, while still leading to an optically pure product. (64)

3.2. Prenyltransferases

Prenyltransferases catalyze the transfer of a prenyl moiety, *e.g.*, a dimethylallyl pyrophosphate (DMAPP), to a variety of substrates: *Trans*- and *cis*-prenyltransferases catalytically attach a prenyl pyrophosphate molecule to an isopentenyl pyrophosphate substrate via an *E*- and *Z*-condensation reaction. (65,66) *Cis*- prenyltransferases are further subdivided into short- (C_{15}), medium- ($C_{50\text{-}55}$) and long-chain ($C_{70\text{-}120}$) *cis*-prenyltransferases. (65,66,67) The corresponding prenylation products, polyterpenes, are important entry points for the biosynthesis of a plethora of natural products and secondary metabolites, *e.g.*, squalene and phytoene. Such metabolites show a wide spectrum of biological activities and are important resources for medicinal research and drug development.

Prenyltransferases also transfer short-chained prenyl donors like geranyl (C_{20}) and farnesyl (C_{15}) pyrophosphate to a conserved cysteine residue in a CaaX motif of proteins and peptides affecting the mechanism of their regulation, localization and function. (68)

Prenyl moieties are also transferred on a multitude of small-size aromatic molecules like phenols, phenolic acids, flavonoids, naphthalenes and indole derivatives. (69) All of which represent promising starting material for lead- and drug-like chemical matter. (70,71,72) In this respect, indole prenyltransferases play a critical role in the biosynthesis of structurally diverse indole alkaloids, which due the prenylation, are further diversified in their respective biosynthetic pathways. (69,73,74) **Figure 6** illustrates the prenylation of L-tryptophan by the prenyltransferase FgaPT2 as starting point of the synthesis of ergotamine. (75)

Indole prenyltransferases – members of the dimethylallyltryptophan synthase (DMATS) superfamily – are soluble proteins and were shown to be suitable for well-yielded overproduction and straightforward purification, rendering their applicability in biotechnological use. (76) They represent one of the most investigated class of prenyltransferases, that show a remarkable flexibility in the acceptance of their aromatic substrate (77,78) and to a smaller extent prenyl substrates. (79) In addition, the regioselectivity of the prenyl transfer reaction strongly depends not only on the indole prenyltransferase itself but also on the distinct combination of aromatic substrate and prenyl moiety donor (**Figure 7**). (80) An interesting conjuncture for *in silico*-driven investigations and protein engineering. A deep understanding of the driving forces behind substrate acceptance, regioselectivity and yield of the prenyl transfer reaction would enable prenyltransferases to be used as an enzyme library – a toolbox – for atom-efficient and green-chemistry-conform hit diversification.

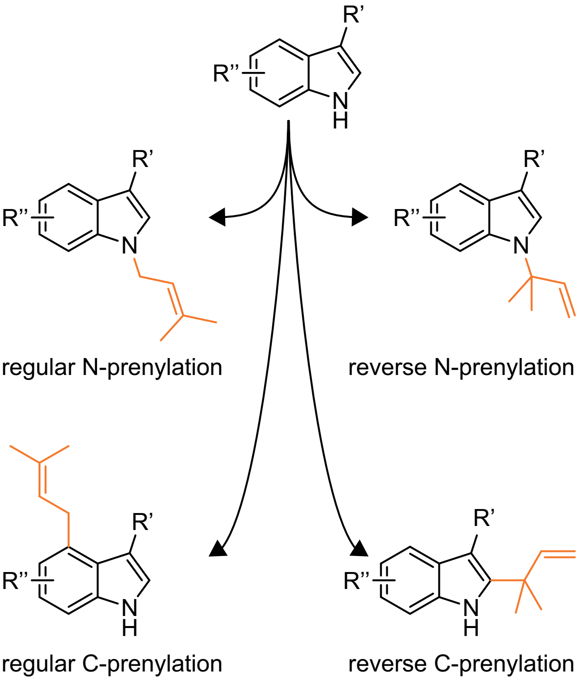


Figure 7: Chemo- and regioselectivity spectrum of indole prenyltransferases.

In this respect, this thesis describes the evolution of prenyltransferases from a well-studied player in the anabolic pathways of secondary metabolites to a systematically organized and algorithmically exploited biocatalysis and compound diversification system. In **Part 2**, I present **PrenDB**, a database of prenyltransferase reactions which – in contrast to databases such as BRENDA (81) or KEGG (82) – although by far not as comprehensive as the former, is more than a listing. By means of a combined fragmentation and subgraph isomorphism approach, it processes reactions conducted by prenyltransferases in an automated manner. Furthermore, it enables the prediction of prenylability of novel compounds based on the reactions deposited in the database.

References

1. Hileman, B. Many Doubt The $800 Million Pharmaceutical Price Tag. *Chemical & Engineering News* **2006,** *84,* 50-51.

2. Leelananda, S. P.; Lindert, S. Computational methods in drug discovery. *Beilstein J. Org. Chem.* **2016,** *12,* 2694–2718.

3. Tollman, P. *A Revolution in R&D: How genomics and genetics are transforming the biopharmaceutical industry;* The Boston Consulting Group: Boston, 2001.

4. Karthick, V.; Nagasundaram, N.; George Priya Doss, C.; Chakraborty, C.; Siva, R.; Lu, A.; Zhang, G.; Zhu, H. Virtual screening of the inhibitors targeting at the viral protein 40 of Ebola virus. *Infectious Diseases of Poverty* **2016,** *5* (12), 1-12.

5. Clark, A. J.; Tiwary, P.; Borrelli, K.; Feng, S.; Miller, E. B.; Abel, R.; Friesner, R. A.; Berne, B. J. Prediction of Protein–Ligand Binding Poses via a Combination of Induced Fit Docking and Metadynamics Simulations. *J. Chem. Theory Comput.* **2016,** *12* (6), 2990–2998.

6. Chao, W.-R.; Yean, D.; Amin, K.; Green, C.; Jong, L. Computer-Aided Rational Drug Design: A Novel Agent (SR13668) Designed to Mimic the Unique Anticancer Mechanisms of Dietary Indole-3-Carbinol to Block Akt Signaling. *J. Med. Chem.* **2007,** *50* (15), 3412–3415.

7. Tran, N.; Van, T.; Nguyen, H.; Le, L. Identification of Novel Compounds against an R294K Substitution of Influenza A (H7N9) Virus Using Ensemble Based Drug Virtual Screening. *Int. J. Med. Sci.* **2015,** *12* (2), 163-176.

8. Talele, T. T.; Khedkar, S. A.; Rigby, A. C. Successful Applications of Computer Aided Drug Discovery: Moving Drugs from Concept to the Clinic. *Current Topics in Medicinal Chemistry* **2010,** *10* (1), 127 - 141.

9. Clark, D. E. What has computer-aided molecular design ever done for drug discovery? *Expert Opinion on Drug Discovery* **2006,** *1* (2), 103-110.

10. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery* **2004,** *3,* 935-949.

11. Lesk, A. *Introduction to Bioinformatics,* 4th ed.; Oxford University Press: Oxford, United Kingdom, 2013.

12. Maithri, G.; Manasa, B.; Vani, S. S.; Narendra, A.; Harshita, T. Computational Drug Design and Molecular Dynamic Studies - A Review. *International Journal of Biomedical Data Mining* **2016,** *6* (1), 1-7.

13. Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K. L. Principles of early drug discovery. *British Journal of Pharmacology* **2011,** *162,* 1239–1249.

14. Hussein, H. A.; Geneix, C.; Petitjean, M.; Borrel, A.; Flatters, D.; Camproux, A.-C. Global vision of druggability issues: applications and perspectives. *Drug Discovery Today* **2017,** *22* (2), 404-415.

15. Volkamer, A.; Kuhn, D.; Grombacher, T.; Rippmann, F.; Rarey, M. Combining Global and Local Measures for Structure-Based Druggability Predictions. *J. Chem. Inf. Model.* **2012,** *52,* 360-372.

16. Brenke, R.; Kozakov, D.; Chuang, G.-Y.; Beglov, D.; Hall, D. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* **2009,** *25* (5), 621-627.

17. Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing

Druggability. *J. Chem. Inf. Model.* **2009,** *49* (2), 377-389.

18. Labute, P.; Santavy, M. SiteFinder-Locating Binding Sites in Protein Structures. https://www.chemcomp.com/journal/sitefind.htm (accessed April 10, 2017).

19. Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012,** *52* (7), 1757-1768.

20. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **2014,** *42,* 1083-1090.

21. Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases. *Nucleic Acids Res.* **2016,** *44* (D1), D1202-D1213.

22. Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014,** *57,* 3186−3204.

23. Bender, A.; Glen, R. B. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004,** *2,* 3204−3218.

24. Kubinyi, H. Similarity and Dissimilarity: A Medicinal Chemist's View. *Perspect. Drug Discovery Des.* **1998,** *9-11,* 225−232.

25. Medina-Franco, J. L.; Maggiora, G. M. Molecular Similarity Analysis. In *Chemoinformatics for Drug Discovery,* 1st ed.; Bajorath, J., Ed.; John Wiley & Sons, Inc, Hoboken, NJ., 2014; pp 343-399.

26. Johnson, M., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity,* 1st ed.; John Wiley & Sons: New York, 1990.

27. Bohacek RS, M. C. G. W. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **1996,** *16* (1), 3-50.

28. Polishchuk, P. G.; Madzhidov, T. I. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput. Aided. Mol. Des.* **2013,** *27,* 675–679.

29. Raymond, J. W.; Willett, P. Maximum Common Subgraph Isomorphism Algorithms for the Matching of Chemical Structures. *J. Comput.-Aided Mol. Des.* **2002,** *16,* 521−533.

30. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010,** *50,* 742−754.

31. Gobbi, A.; Poppinger, D. Genetic Optimization of Combinatorial Libraries. *Biotechnol. Bioeng.* **1998,** *61,* 47-54.

32. Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002,** *42* (6), 1273–1280.

33. Renner, S.; Schneider, G. Scaffold-Hopping Potential of Ligand-Based Similarity Concepts. *ChemMedChem* **2006,** *1,* 181-185.

34. Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold Hopping Using Two-Dimensional Fingerprints: True Potential, Black Magic, or a Hopeless Endeavor? Guidelines for Virtual Screening. *J. Med. Chem.* **2010,** *53,* 5707-5715.

35. Good, A. C.; Richards, W. G. Explicit Calculation of 3D Molecular Similarity. *Perspect. Drug Discovery Des.* **1998,** *9-11,* 321-338.

36. Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein−Protein Interaction. *J. Med. Chem.* **2005,** *48,* 1489−1495.

37. Berman, H. M. The protein data bank. *Nucleic Acids Res.* **2000,** *28,* 235-242.

38. Meng, X. Y.; Zhang, H. X.; Mezei, M.; Cui, M. Molecular docking: A powerful approach for structure-based drug discovery. *Curr. Comput. Aided Drug Des.* **2011,** *7,* 146-157.

39. López-Vallejo, F.; Caulfield, T.; Martínez-Mayorga, K.; Giulianotti, M. A.; Houghten, R. A.; Nefzi, A.; Medina-Franco, J. L. Integrating virtual screening and combinatorial chemistry for accelerated drug discovery. *Comb. Chem. High Throughput Screen.* **2011,** *14,* 475-487.

40. Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein-ligand docking: Current status and future challenges. *Proteins Struct. Funct. Bioinform.* **2006,** *65,* 15-26.

41. Gorelik, B.; Goldblum, A. High quality binding modes in docking ligands to proteins. *Proteins Struct. Funct. Bioinform.* **2008,** *71,* 1373-1386.

42. McGann, M. FRED and HYBRID docking performance on standardized datasets. *J. Comput. Aided Mol. Des.* **2012,** *26,* 897-906.

43. Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.* **2001,** *15,* 411-428.

44. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shaw, D. E.; Shelley, M.; al., e. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004,** *47,* 1739-1749.

45. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996,** *261,* 470-489.

46. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997,** *267,* 727-748.

47. Corbeil, C. R.; Williams, C. I.; Labute, P. Variability in docking success rates due to dataset preparation. *J. Comput. Aided Mol. Des.* **2012,** *26,* 775-786.

48. Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed automated docking of flexible ligands to proteins: Parallel applications of AutoDock 2.4. *J. Comput. Aided Mol. Des.* **1996,** *10,* 293-304.

49. Huang, S. Y.; Zou, X. Advances and challenges in protein-ligand docking. *Int. J. Mol. Sci.* **2010,** *11,* 3016-3034.

50. Jain, A. N. Scoring functions for protein-ligand docking. *Curr. Protein Pept. Sci.* **2006,** *7,* 407-420.

51. Roughley, S. D.; Jordan, A. M. The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.* **2011,** *54,* 3451-3479.

52. Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012,** *52,* 2864-2875.

53. Meyers, J.; Carter, M.; Mok, N. Y.; Brown, N. On the origins of three-dimensionality in drug-like molecules. *Future Med. Chem.* **2016,** *8* (14), 1753-1767.

54. Bommarius, A. S. Biocatalysis: A Status Report. *Annu. Rev. Chem. Biomol. Eng.* **2015,** *6,* 319-345.

55. Fersht, A. *Structure adn mechanism in protein science: A guide to enzyme catalysis and protein folding,* 1st ed.; W. H. Freeman and Company: New York, 1999.

56. Anastas, P. T.; Warner, J. C. *Green Chemistry: Theory and Practice,* 1st ed.; Oxford University Press, USA: New York, 1998.

57. Reetz, M. T. Laboratory evolution of stereoselective enzymes: a prolific source of catalysts for asymmetric reactions. *Angew. Chem. Int. Ed.* **2011,** *50,* 138–174.

58. Reetz, M. T.; Carballeira, J. D. Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nat. Protoc.* **2007,** *2,* 891–903.

59. Reetz, M. T.; Kahakeaw, D.; Lohmer, R. Addressing the numbers problem in directed evolution. *ChemBioChem* **2008,** *9,* 1797-1804.

60. Reetz, M. T.; Bocola, M.; Carballeira, J. D.; Zha, D. X.; Vogel, A. Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test. *Angew. Chem. Int. Ed.* **2005,** *44,* 4192-4196.

61. Oikawa, H.; Kobayashi, T.; Katayama, K.; Suzuki, Y.; A., I. Total synthesis of (−)-solanapyrone A via enzymatic Diels-Alder reaction of prosolanapyrone. *J. Organ. Chem.* **1998,** *63,* 8748-8756.

62. Reetz, M. T.; Mondiere, R.; Carballeira, J. D. Enzyme promiscuity: first protein-catalyzed Morita-Baylis-Hillman reaction. *Tetrahedron Lett.* **2007,** *48,* 1679-1681.

63. Rothlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Baker, D. Kemp elimination catalysts by computational enzyme design. *Nature* **2008,** *453,* 190-195.

64. Desai, A. A. Sitagliptin Manufacture: A Compelling Tale of Green Chemistry, Process Intensification, and Industrial Asymmetric Catalysis. *Angew. Chem. Int. Ed.* **2011,** *50,* 1974-1976.

65. Kharel, Y.; Takahashi, S.; Yamashita, S.; Koyama, T. Manipulation of prenyl chain length determination mechanism of cis-prenyltransferases. *FEBS J* **2006,** *273,* 647-657.

66. Lu, Y. P.; Liu, H. G.; Liang, P. H. Different reaction mechanisms for cis and trans-prenyltransferases. *Biochem. Biophys. Res. Commun.* **2009,** *379,* 351-355.

67. Takahashi S, K. T. Structure and function of cis-prenyl chain elongating enzymes. *Chem. Rec.* **2006,** *6,* 194-205.

68. Perez-Sala, D. Protein isoprenylation in biology and disease: general overview and perspectives from studies with genetically engineered animals. *Front. Biosci.* **2007,** *12,* 4456-4472.

69. Heide, L. Prenyl transfer to aromatic substrates: genetics and enzymology. *Curr. Opin. Chem. Biol.* **2009,** *13,* 171-179.

70. Liu, A. H.; Liu, D. Q.; Liang, T. J.; Yu, X. Q.; Feng, M. T.; Yao, L. G.; Fang, Y.; Wang, B.; Feng, L. H.; Zhang, M. X.; Mao, S. C. Caulerprenylols A and B, two rare antifungal prenylated para-xylenes from the green alga Caulerpa racemosa. *Bioorg. Med. Chem. Lett.* **2013,** *23,* 2491-2494.

71. Oya, A.; Tanaka, N.; Kusama, T.; Kim, S. Y.; Hayashi, S.; Kojoma, M.; Hishida, A.; Kawahara, N.; Sakai, K.; Gonoi, T.; and Kobayashi, J. Prenylated benzophenones from Triadenum japonicum. *J. Nat. Prod.* **2015,** *78,* 258-264.

72. Li, S.-M. Prenylated indole derivatives from fungi: structure diversity, biological activities, biosynthesis and chemoenzymatic synthesis. *Nat. Prod. Rep.* **2010,** *27,* 57-78.

73. Winkelblech, J.; Fan, A.; Li, S.-M. Prenyltransferases as key enzymes in primary and secondary metabolism. *Appl. Microbiol. Biotechnol.* **2015,** *99,* 7379-7397.

74. Williams, R. M.; Stocking, E. M.; Sanz-Cervera, J. F. Biosynthesis of prenylated alkaloids derived from tryptophan. *Topics Curr. Chem.* **2000,** *209,* 97-173.

75. Gerhards, N.; Neubauer, L.; Tudzynski, P.; Li, S.-M. Biosynthetic Pathways of Ergot Alkaloids. *Toxins* **2014,** *6,* 3281-3295.

76. Li, S.-M. Applications of dimethylallyltryptophan synthases and other indole prenyltransferases for structural modification of natural products. *Appl. Microbiol. Biotechnol.* **2009,** *84,* 631-639.

77. Kremer, A.; Li, S.-M. Potential of a 7-dimethylallyltryptophan synthase as a tool for production of prenylated indole derivatives. *Appl. Microbiol. Biotechnol.* **2008,** *79,* 51-961.

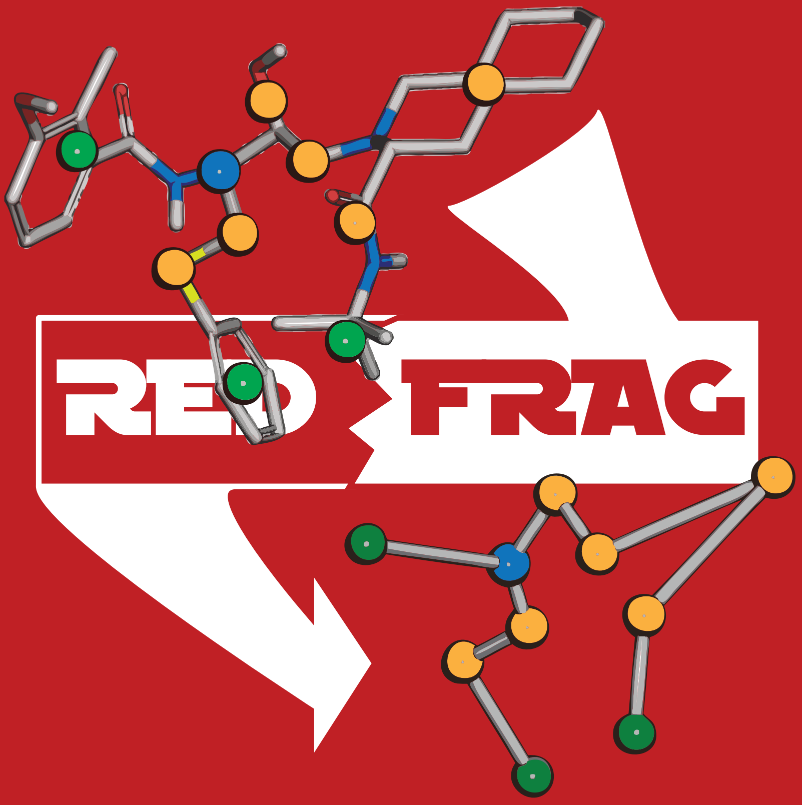
78. Steffan, N.; Unsöld, I. A.; Li, S.-M. Chemoenzymatic synthesis of prenylated indole derivatives by using a 4-dimethylallyltryptophan synthase from Aspergillus fumigatus. *Chembiochem* **2007,** *8,* 1298-1307.

79. Grundmann, A.; Kuznetsova, T.; Afiyatullov, S. S.; Li, S.-M. FtmPT2, an N-prenyltransferase from Aspergillus fumigatus, catalyses the last step in the biosynthesis of fumitremorgin B. *Chembiochem* **2008,** *9,* 2059-2063.

80. Winkelblech, J.; Liebhold, M.; Gunera, J.; Xie, X.; Kolb, P.; Li, S.-M. Tryptophan C5-, C6- and C7-Prenylating Enzymes Displaying a Preference for C-6 of the Indole Ring in the Presence of Unnatural Dimethylallyl Diphosphate Analogues. *Adv. Synth. Catal.* **2015,** *357,* 975-986.

81. Schomburg, I.; Chang, A.; Ebeling, C.; Gremse, M.; Heldt, C.; Huhn, G.; Schomburg, D. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.* **2004,** *32,* D431–D433.

82. Kanehisa, M.; Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000,** *28,* 27-30.



Journal of Computational Chemistry **2015**, *36*, 1597-1608.

Part 1

Part 1 captures my first full article in which Peter and I present a new method for two-dimensional molecular similarity calculation. RedFrag combines two popular approaches in the field: Similarity calculation via holistic fingerprints and molecular pattern matching via graph isomorphism. Our approach exploits the idea of molecules being a construct of smaller chemical entities, fragments, simultaneously bypassing the necessity of defining sets of substructures, moieties or groups. The utilization of only a few – or even only one – fragmentation rules allows to look on molecule as a combination of fragments which are represented by a potentially infinite number of hues in the color space spanned by the underlying fingerprints. This enables RedFrag to discover similar molecules on their composition level, allows for fragmental diversity but still retains the molecular topology.

The author list is the following (by contribution order): Gunera J, Kolb P. I was responsible for the development of RedFrag's algorithm: The design and layout of the underlying code structure and the programming. I performed the retrospective validation study and parameter optimization as well as the prospective virtual screen for putative endothiapepsin ligands.

Fragment-based similarity searching with infinite color space

Jakub Gunera[1] and Peter Kolb[1]

[1]Department of Pharmaceutical Chemistry, Philipps-University, Marburg, Hesse, 35032, Germany

Keywords:

molecule fragmentation, reduced graphs, fragment-based similarity, scaffold hopping, endothiapepsin

Abstract

Fragment-based searching and abstract representation of molecular features through reduced graphs have separately been used for virtual screening. Here, we combine these two approaches and apply the algorithm RedFrag to virtual screens retrospectively and prospectively. It uses a new type of reduced graph that does not suffer from information loss during its construction and bypasses the necessity of feature definitions. Built upon chemical epitopes resulting from molecule fragmentation, the reduced graph embodies physico-chemical and 2D-structural properties of a molecule. Reduced graphs are compared with a continuous-similarity-distance-driven maximal common subgraph algorithm, which calculates similarity at the fragmental and topological levels. The performance of the algorithm is evaluated by retrieval experiments utilizing pre-compiled validation sets. By predicting and experimentally testing ligands for endothiapepsin, a challenging model protease, the method is assessed in a prospective setting. Here, we identified five novel ligands with affinities as low as 2.08 µM.

1. Introduction

In the past decades, ligand-similarity-based methods for virtual screening of large molecular databases have increased in importance. This is due to the steady and concomitant growth of databases holding small organic compounds (1) and experimentally validated ligands for a plethora of potential protein targets. (2) Similarity calculations and searches are based upon the similarity principle in medicinal chemistry, which states that similar molecules are likely to have similar biological effects. (3) While this certainly seems intuitive and has been shown by, *e.g.*, Martin *et al.*, (4) it immediately leads to the question of how to define similarity between molecules. Historically, this question has been answered by developments based on notions such as fingerprints and distances or similarity coefficients in order to quantify the extent to which two molecules can be considered similar. (5,6,7,8) Such classical concepts of similarity tend to focus on the global resemblance of a pair of compounds.

The scope of similarity determination ranges from 2D methods such as substructure (9) and pharmacophore searching (10) to 3D methods such as molecular shape overlay (11) and the conceptually similar topomer similarity. (12,13) A common theme of the most frequently used approaches is the encoding of molecules via descriptors representing their structural and/or physico-chemical properties. In principle, such fingerprints can be constructed at different levels of granularity, ranging from atom counts to the occurrence of certain substructures. Fingerprints are usually fast and therefore attractive for screening large libraries. However, they can suffer from a drawback which is ingrained in their very nature: Since they focus on similarity, they will of course yield molecules that can be very similar to each other. This is especially bothersome when scaffold hops are desired (14), which is often the case when investigating well-researched targets. Thus, fingerprint methods that have a slightly different view of what similarity means and that do not focus on the precise atom arrangement are desirable. Of course, 3D pharmacophore searches or docking could be employed to that end, too, but they have their own drawbacks when it comes to the generation of conformations and speed.

At the low-resolution end of the spectrum of similarity methods, feature trees have been developed. (20,21) They reduce functional groups of a molecule to nodes of different type, simplifying the graph of a molecule by going from individual atoms as nodes to cycle-free trees with nodes as higher-order features. Commonly used implementations exploit several levels of graph reduction. (22) These are then combined with a pre-defined, limited number of colors (or labels) that represent different chemical features. (23)

Depending on the underlying granularity of node detection and the color space applied during the node coloring process, information is lost and thus becomes unavailable for comparison procedures. Moreover, by using a finite and pre-defined set of colors, the user will unavoidably exert an influence on the amount of novelty that can be expected.

In an effort to further develop the concept of ligand similarity based on reduced graphs, we have developed a fragment-based approach that uses an infinite, *i.e.*, label free, color space, while retaining the efficiency of the original concept. We keep another important notion of feature trees, which is the straightforwardness with which the features (in our case: fragments) can be recognized as sensible chemical entities by a chemist. To cater to different tastes of different chemists, three fragmentation rules have been implemented, expandable by the user. By combining these concepts with a reduced graph similarity approach developed earlier by, *e.g.*, Gillet *et al.* (22) and Barker *et al.* (24), we also achieve independence from the way in which the fragments are connected to each other. This enables the algorithm to explore a broader range of similarities and also introduces the potential for scaffold hops.

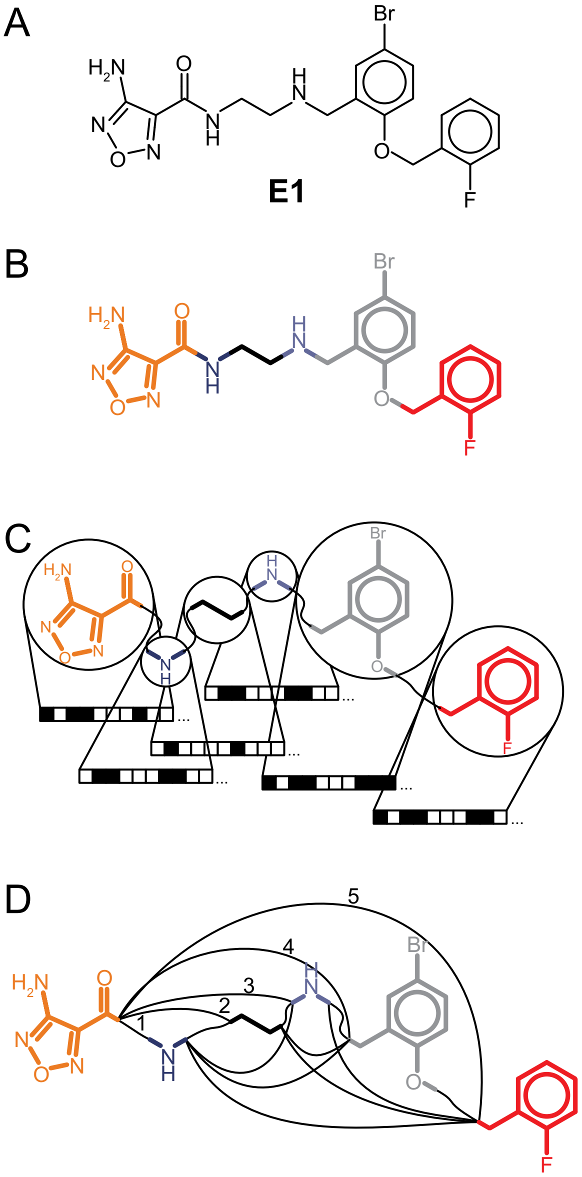


Figure 1: Fragmentation of an example molecule (**E1**) from the MUV S1P1R sample: **A** Molecular structure of compound **E1**. **B** Fragment pattern after RECAP cleavage. Every fragment is highlighted in a different color. **C** After RFGraph-characterization, connectivity perception and compression, the fragments are assigned to fingerprint objects. **D** Complete RFGraph ready for entering the scoring algorithm.

A scaffold hop is encountered when a compound is similarly active, but shows significant structural differences in either topological arrangement of chemical features (epitopes) or the exchange of core or peripheral structural motifs compared to a reference compound. (25,26,27,28) The capability of a similarity method to identify scaffold hops has recently become a very popular topic in the community. (29,30) It is important to not only predict close analogs or homologs of known active compounds but to also discover novel structural themes. Bioisosterism, *i.e.*, the replacement of functional groups or larger chemical epitopes by physico-chemically similar groups, is a common way to achieve structural

Table 1: Seven fingerprints used in this study for encoding chemical epitopes.

Full name	Abbreviation	Description
Topological Fingerprint	topo	identifies and hashes topological paths in the molecule, an RDKit reimplementation of the Daylight fingerprint
MACCS Keys	maccs	SMARTS-based implementation of the 166 public MACCS Keys
Atom Pairs Fingerprint	apairs	based on atom types derived from each chemical environment and the topological distance between each pair of atoms (15)
Topological Torsions Fingerprint	torsions	fingerprint consists of hashed quartets of atoms typed by their chemical environment (16)
Morgan/Circular Fingerprint	ecfp	based on circular neighborhood for each atom within a given radius (17)
Morgan/Circular Fingerprint	fcfp	its feature-based variant where SMARTS encoded structural properties are taken into account (18)
DAIM Fingerprint	daim	based on atom and property counts as described by Kolb and Caflisch (19)

diversity while retaining bioactivity. In order to assess the usefulness of looking at chemical similarity in a fragment-based way, we have tested the algorithm on three levels, which is mirrored in the structure of this manuscript. First, we applied it to a non-trivial test set for ligand-based methods, the Maximum Unbiased Validation data sets (MUV) by Rohrer *et al.* (31); Second, we evaluated its scaffold enrichment performance based on a scaffold distance defined by Li *et al.* (32); Third, we predicted and experimentally verified novel ligands for endothiapepsin, an aspartic protease, which commonly is regarded as a challenging target in terms of the identification of small-molecule binders.

2. Materials and Methods

2.1. RedFrag package

Our strategy for label-free fragment-based similarity searching builds on a combination of classical graph isomorphism-driven and fingerprint-supported similarity calculations to retain benefits from fragment-based drug discovery on the one hand and overcome the necessity of pre-defined fragment dictionaries on the other. REDuced graph FRAGment based similarity search tool (RedFrag) has been developed as a python package designed to i) read molecules from the most common file formats; ii) decompose them according to fragmentation rules such as RECAP (33), BRICS (34), DAIM (19) and user-defined patterns; and, iii) abstract these molecules to reduced fragment-based graphs (RFGraphs) and evaluate a similarity score between pairs of RFGraphs using seven different fingerprints. All programming involving molecule file reading, fingerprint generation and similarity calculation was done using the RDKit open source toolkit for chemoinformatics. (35) RFGraph creation and manipulation utilizes the NetworkX package. (36) In more detail, we implemented and tested the seven fingerprints listed in **Table 1**. This allowed us to unambiguously assess the value added by fragmentation.

The way in which a molecule is fragmented affects the resulting reduced graph in terms of size and complexity. Instead of relying on pre-defined fragment or functional group definitions, bonds of a molecule are cleaved to yield chemically reasonable fragments in our approach. A fragment is thus defined as a set of atoms connected by unbreakable bonds, and physico-chemical descriptors can readily be derived for such fragments. The topology of the parent molecule is preserved by retaining the connectivity of the individual fragments and converting detected fragments into nodes of the RFGraph. Importantly, each of these nodes of a reduced graph contains the complete chemical identity of the fragment it represents in the form of a SMILES (37) string and the corresponding fingerprint.

Thus, the color of the label of each chemical epitope comes from a color space with an infinite number of hues rather than a pre-defined and thus limited color dictionary. To avoid noise from nodes representing very small fragments, *e.g.*, methylene groups, a subsequent graph compression step removes such adjacent nodes from the graph. During the last step of graph reduction, the compressed RFGraphs are transformed into complete graphs by connecting all nodes of the graph with each other. Edges are weighted with the topological distance between pairs of nodes, resulting in a labeled and weighted graph (**Figure 1A-D**).

2.2. Reduced-graph fragment-based similarity search

Scoring the similarity of two molecules **R** and **T** (for reference and target molecule, *e.g.*, **E1** and **E2** in **Figure 2**) in our approach is subdivided into four distinct steps: First, the molecules are reduced to their complete RFGraphs, $R(V_R, E_R)$ and $T(V_T, E_T)$, V and E being the sets of nodes and edges, respectively. Second, the correspondence graph $CG_{RT}(V_R \times V_T)$, a product of both sets of RFGraph nodes, is constructed. Each node of the correspondence graph $[n_i, n_j]$ consists of a matched pair of a reference and target RFGraph node. Such a match occurs only when the Tanimoto similarity between the nodes exceeds a certain threshold t, *i.e.*, the node colors are within a given color interval: For $n_i \in V_R$ and $n_j \in V_T$, $[n_i, n_j] \in V_{CG_{RT}}$ if $T_D(f_i, f_j) \geq t$, with T_D being the Tanimoto similarity (6,38), n_i, n_j nodes (fragments) of reference and target RFGraph, f_i, f_j being their corresponding fingerprints, and $V_{CG_{RT}}$ the node-set of the correspondence graph, respectively (**Figure 2A**). Third, the nodes of matched fragment pairs are connected. An edge between two nodes $[n_i, n_j]$ and $[n_k, n_l]$ within the correspondence graph exists if the topological distance $w(n_i, n_k)$ between the two nodes corresponding to the reference graph, equals the topological distance $w(n_j, n_l)$ between the two corresponding nodes

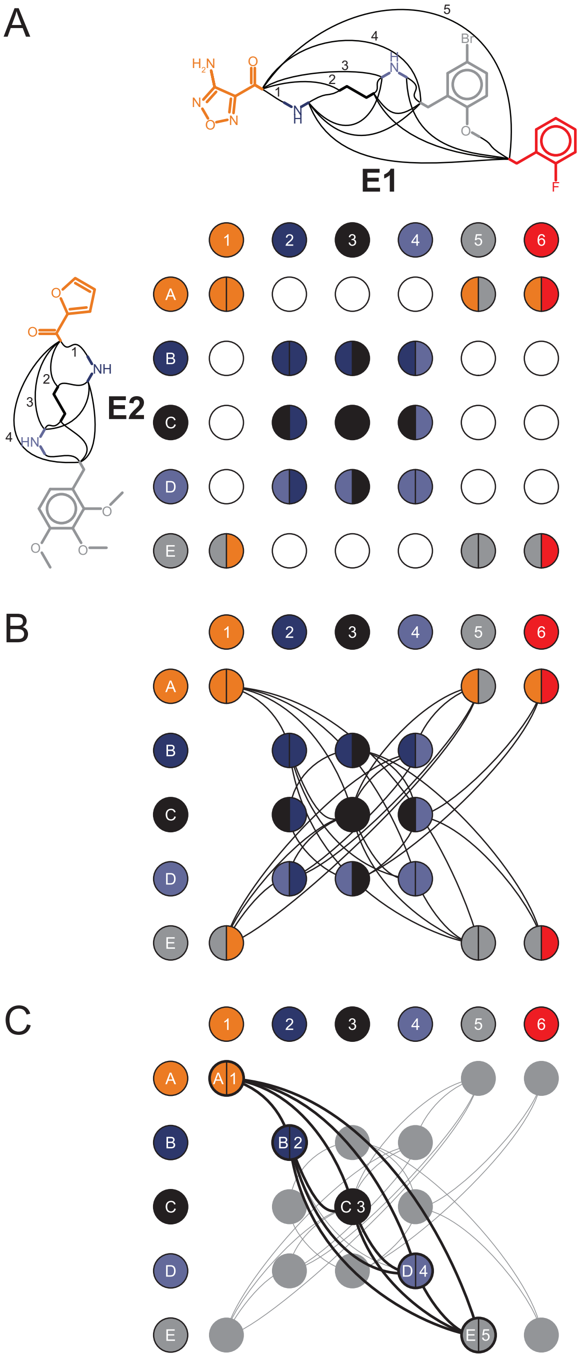


Figure 2: Schematic of the RedFrag algorithm: **A** Correspondence graph creation based on the reduced graphs of compounds **E1** and **E2**. **B** Nodes of the correspondence graph are connected according to the condition mentioned in the text. **C** Highlighted in black is the clique of maximum size of the correspondence graph. It corresponds to the maximum common subgraph of the underlying RFGraphs within a given similarity threshold.

within the target RFGraph: Edge $\{[n_i, n_j], [n_k, n_l]\} \in E_{CG_{RT}}$ if $w(n_i, n_k) = w(n_j, n_l)$ (**Figure 2B**), $E_{CG_{RT}}$ being the edge-set of the correspondence graph. If the correspondence graph exists, a maximum clique detection algorithm is invoked, where the maximum clique MC_{RT} corresponds to the maximum common subgraph of the reference and target graphs (**Figure 2C**). (39,40,41) By deconvolution of the maximum common subgraphs back onto the reference and target molecules, the maximum overlapping set of common substructures (here fragments) can be obtained. (42) In the final step, the resulting clique is scored per scoring function (eq 1), which consists of two parts: i) the graph similarity G_T, which reflects the overall topological congruency of the two input reduced graphs (eq 2) and ii) the fragment similarity F_T, where the quality of all fragment-to-fragment matchings is taken into account (eq 3). In order to be able to tune the relative importance of the two terms G_T and F_T with respect to each other, we introduce two weighting parameters G_C and F_C. This allows to emphasize either graph congruency or chemical similarity in the final score:

$$S(R,T) = \frac{G_C \cdot G_T(R,T) + F_C \cdot F_T(R,T)}{G_C + F_C} \quad (1)$$

$$G_T(R,T) = \frac{|MC_{RT}|}{|R| + |T| - |MC_{RT}|} \quad (2)$$

$$F_T(R,T) = \frac{1}{|MC_{RT}|} \sum_{i=1}^{|MC_{RT}|} T_D([f_R, f_T]_i) \quad (3)$$

Where $|MC_{RT}|$, $|R|$ and $|T|$ are the sizes of the maximum clique, the reference and the target RFGraph, respectively. T_D is the Tanimoto distance between a fragment of graph R and graph T based on the corresponding fingerprints f_R and f_T, $[f_R, f_T]_i$ being a fingerprint-pair corresponding to the ith composite node of the maximum clique MC_{RT}.

2.3. Validation data set

We validated and optimized our method against the Maximum Unbiased Validation data sets (MUV). (31) The MUV contains 17 activity classes, and each class encompasses 30 actives that have been determined via a highly reliable experimental method. Each set of actives is accompanied by 15,000 decoys (see **Table S1**).

2.4. Validation procedure

We investigated the capability of RedFrag to separate active molecules from decoys when passing a single active molecule as reference. First, one active was selected randomly from the set of 30 actives of an activity class. Then, the remaining actives were mixed into the set of 15,000 decoys and the fragment-based reduced-graph similarity search was invoked. The resulting ranked list of target molecules was visualized by plotting receiver operating characteristic (ROC) curves. As a measure of retrieval performance, the area under the curve (AUC) value was calculated. As we chose to investigate a total of three fragmentation patterns (BRICS, RECAP, DAIM), seven fingerprints (apairs, topo, maccs, torsions, ecfp, fcfp, daim, **Table 1**), nine different similarity thresholds (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9) and seven distinct combinations of scoring function coefficients G_C and F_C (0.5-3.5, 1.0-3.0, 1.5-2.5, 2.0-2.0, 2.5-1.5, 3.0-1.0, 3.5-0.5), a total of 674,730 (30·17·3·7·9·7) calculations had to be run and analyzed.

2.5. Analysis of results

In order to investigate the effectiveness of our algorithm and simultaneously handle the large number of individual calculations, we decided to average the outcomes of the retrieval experiments at the level of a single MUV activity class. For each combination of fingerprint, similarity threshold and scoring function coefficients, the mean and median AUC value of all 17 activity classes was calculated, representing the overall performance of the method with the respective parameters. RedFrag's potential to perform scaffold hops was assessed by first calculating an average scaffold distance (32) between each active compound and the remaining actives within an activity class. Based on the average scaffold distance, we defined, for each activity class, a set of distant scaffolds, *i.e.*, actives with an average scaffold distance to all other actives

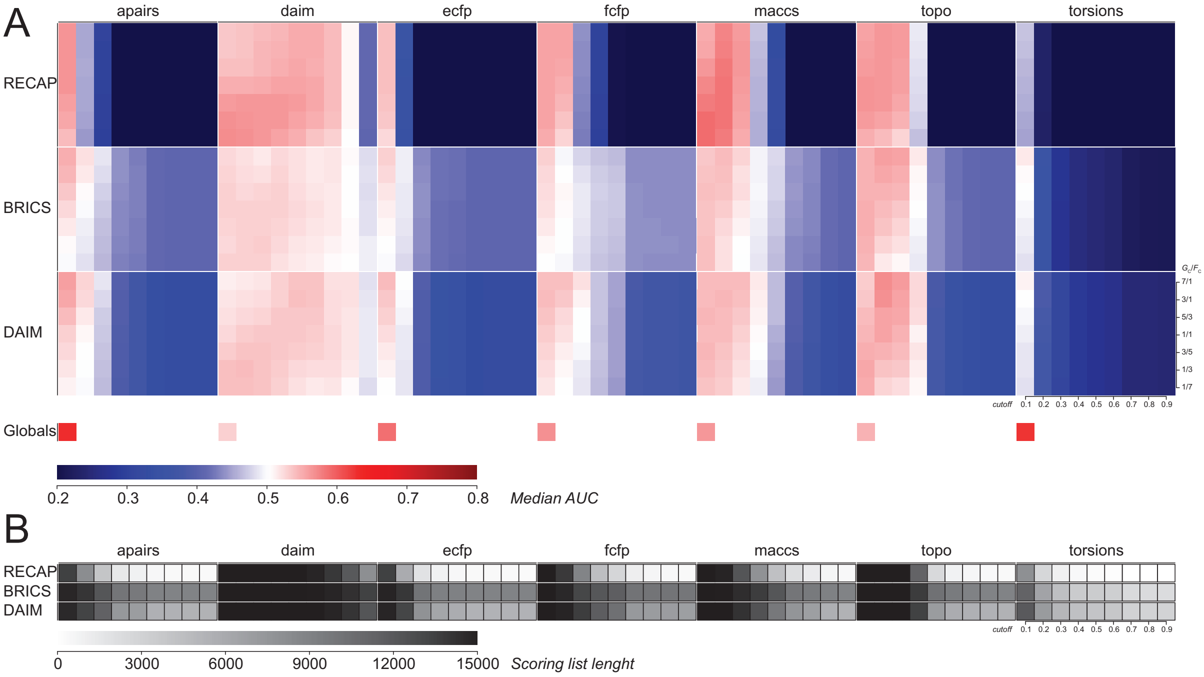


Figure 3: The effectiveness of the underlying method is shown for three fragmentation patterns: RECAP, BRICS and DAIM and seven fingerprints: **A** Each block consists of 63 (7 fingerprints and 9 similarity thresholds) averaged calculation series represented as colored squares. The more reddish a square, the higher the AUC value averaged over all 17 activity classes and 30 retrieval experiments each, thus the better the performance. Blueish squares correspond to worse than random averaged performance. Each square corresponds to 510 retrieval experiments (30 actives in each of the 17 activity classes). **B** In the same arrangement as in **A**, averaged lengths of scoring lists are represented as shaded squares. A darker square corresponds to a longer scoring list, thus to a larger number of successful scoring events.

greater than 0.56 (this number being the average distance within a library of distinct scaffolds as described in ref. (32)). Subsequently, we analyzed the retrieval experiments with respect to the number of distant scaffolds successfully enriched in the top 10 % of the scoring list. Furthermore, we analyzed the similarity between our virtual hits and the query molecules we used for virtual screening based on ECFP4 (17) fingerprints.

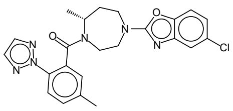
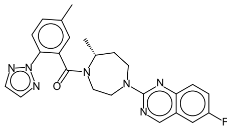
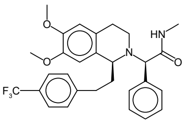
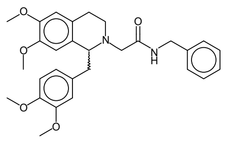
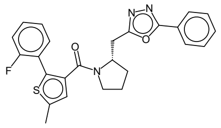
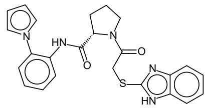
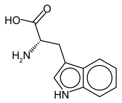
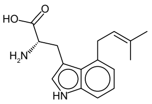
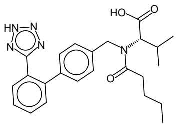
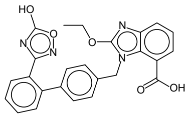
2.6. Virtual screen of the ZINC database

In order to prospectively validate our algorithm, we wanted to find endothiapepsin binders in a screen of the DrugsNow subset of the ZINC database. (1,43) As queries, we chose 67 compounds with known activity towards endothiapepsin from three different sources: i) the ChEMBL database (44) (11 compounds), ii) a recent publication by Mondal *et al.* (45) (7 compounds) and 49 compounds from an SAR study of aspartic proteases. (46) Each of these 67 compounds was used in a RedFrag search and scored with equation 1. For comparison, we used two distinct sets of parameters: i) RECAP|maccs|0.1|0.5-3.5 and ii) RECAP|maccs|0.2|3.5-0.5 reflecting the two best-performing parameter sets from the retrospective study and including either emphasis on fingerprint similarity or graph congruency. For each run, the top 10,000 scores were further processed by sphere exclusion clustering (47) in order to reduce the number of compounds to be visually inspected.

2.7. Endothiapepsin functional assay

To investigate the biological activity of the selected compounds, we performed inhibition studies based on a fluorometric assay adapted from the HIV-Protease assay described by Toth and Marshall (48). The assays were carried out as described earlier in ref. (45). Stock solutions (50 mM in DMSO) were prepared for all molecules subjected to the assay. As substrate, Abz-Thr-Ile-Nle-p-nitro-Phe-Gln-Arg-NH2 (Bachem, Basel) was used. The assay was performed in flat-bottom 96-well microplates (Greiner Bio-One, Frickenhausen) and

Table 2: Comparison of scores of example pairs of molecules.[a]

	Reference structure		Target structure	ECFP4[b]	PF[b]	LINGO[c]	MACCS[c]	RedFrag[d]
R1[e]		**T1**[e]		0.545	0.928	0.692	0.848	0.930
R2[f]		**T2**[f]		0.432	0.640	0.285	0.800	0.728
R3[g]		**T3**[g]		0.223	0.597	0.270	0.606	0.337
R4[h]		**T4**[h]		0.565	0.625	0.636	0.820	0.860
R5[i]		**T5**[i]		0.294	0.452	0.273	0.546	0.635

[a]Similarities were calculated as Tanimoto coefficients.
[b]Implementation based on the ChemAxon API for Java (PF: Pharmacophore fingerprint). (50)
[c]Implementation based on the OpenEye OEChem and OEGraphSim TK API for Python. (51)
[d]Calculation based on the top performing RECAP|maccs|0.1|0.5-3.5 parameter set.
[e]Suvorexant (**R1**), dual orexin receptor antagonist approved insomnia drug, and its analog **T1**.
[f]Almorexant (**R2**), dual orexin receptor antagonist, and its pre-optimized scaffold **T2**.
[g]Orexin 1 receptor antagonist **R3** developed by GSK and orexin 2 receptor selective analog **T3** developed by Merck.
[h]L-Tryptophan (**R4**) and its prenylated analog **T4**.

fluorescence resulting from cleaved product was recorded in a Synergy Mx microplate reader at an extinction wavelength of 337 nm and an emission wavelength of 414 nm. The K_m of the substrate towards endothiapepsin was determined to be 1.6 µM. (49) The assay buffer (0.1 M sodium acetate, pH 4.6, 0.001 % Tween 20) was premixed with substrate and potential inhibitors, while endothiapepsin was added directly before measurements commenced. The final reaction volume was 200 µL, containing 0.4 nM endothiapepsin, 1.8 µM substrate and 500 µM compound. In the same way, blanks were prepared using DMSO instead of compound stock solution. Each compound was measured in duplicate, results reported herein are the average of both measurements.

3. Results

3.1. Retrospective study

3.1.1. Parameter optimization using the MUV data sets

The effectiveness of our fragment-based reduced graph approach RedFrag at retrieving active compounds from mixed active-decoy sets was investigated using the MUV data sets (17 activity classes, 30 experimentally validated active compounds and 15,000 decoys per class). In a first step, the retrieval rate was used to assess the influence of the few high-level choices and parameters that are available, *viz.*, fragmentation patterns and type of fingerprint used, as well as the similarity threshold and scoring function coefficients. This resulted in 674,730 retrieval experiments, which are shown in **Figure 3A** and **B** in a compressed way. Each colored square corresponds to the averaged AUC value originating from 510 retrieval experiments partitioned over 17 activity classes. The AUC value is reflected in the color of a square: deep red squares reflect better-than-random and blue squares worse-than-random performance, respectively. White squares show random performance. The highest averaged AUC value over all activity classes, were systematically achieved when employing the RECAP fragmentation rules. The numerically best performance coincided with a low G_C vs. F_C ratio for daim and maccs fingerprints and high G_C vs. F_C ratio for apairs, ecfp, fcfp and topo fingerprints, respectively. The ratio of the scoring function coefficients emphasizes either the topological congruency (eq 2) of reference and target graph or the Tanimoto similarity of their fragments (eq 3). In order to relate to a specific set of parameters and RedFrag's corresponding performance, we will use a tag consisting of the fragmentation pattern, fingerprint, similarity cutoff, scoring function coefficients and the average AUC value in this order and separated by vertical bars and a double colon, *e.g.*, RECAP|maccs|0.1|0.5-3.5::0.59 corresponds to a retrieval experiment performed with the RECAP fragmentation rules, the fragments encoded with MACCS Keys fingerprint, a relaxed similarity criterion for fragment-to-fragment matching, and a scoring function with an emphasis on the fingerprint term, resulting in a performance of 0.59.

3.1.2. Comparison of fingerprint metrics

In order to facilitate understanding of the RedFrag score prior to a detailed analysis of its retrospective performance, we compare it to commonly used fingerprint metrics here. **Table 2** shows five example molecule pairs and their corresponding similarities derived from their fingerprint representations. The molecules were selected per the high prominence of their targets and their holistic and fragment or epitope similarity. Although, despite numerical

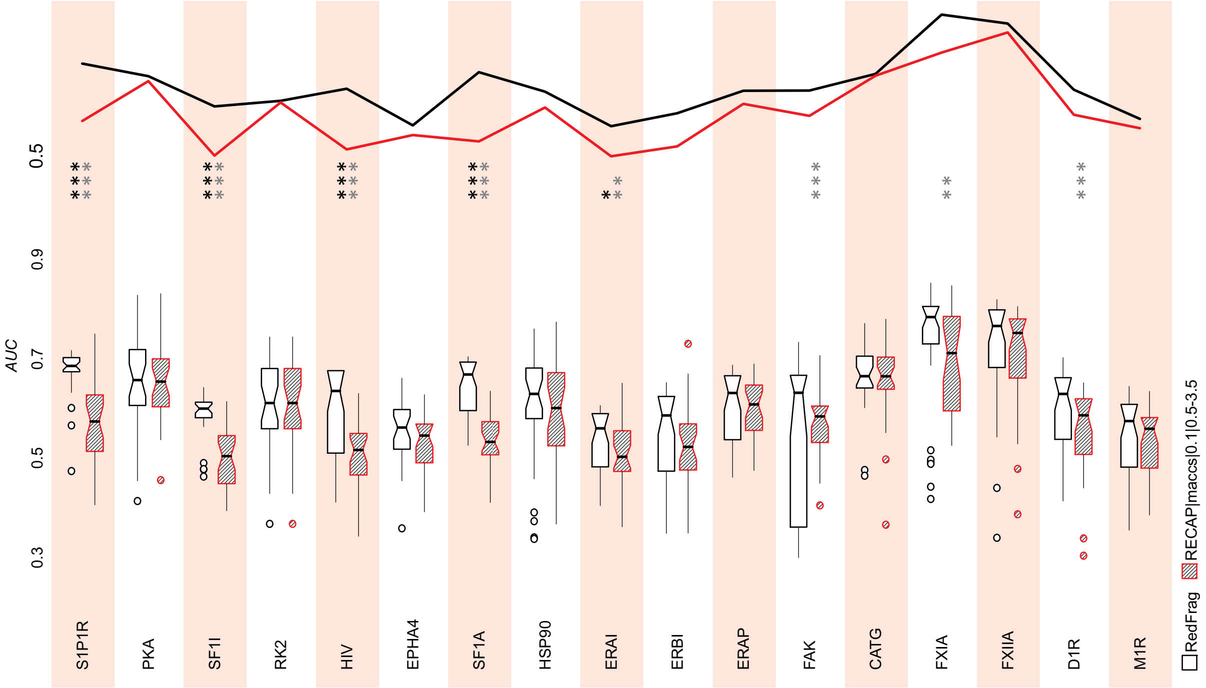


Figure 4: Shown are distributions of AUC values for each activity class as box-plot representations. Each AUC distribution corresponds to 30 retrieval experiments. Black data correspond to the OOPS while red data show the fixed parameter set. Solid curves emphasize the trend of the medians. Black and gray asterisks show the level of statistical significance of the AUC distributions within an activity class as calculated by the t-test and Kolmogorov-Smirnov-test, respectively. (∗∗∗→ $p < 0.01$, ∗∗→ $p < 0.05$, ∗→ $p < 0.10$).

differences, the trend of the scores changes in a unidirectional manner, each representation results in different similarity estimates and varying similarity differences, further illustrating the vagueness of what similarity is (see also **Figure S1**).

3.1.3. Descriptive analysis of Figure 3

As can be seen in **Figure 3A**, our algorithm is sensitive to both the choice of fingerprint as well as the set of fragmentation rules. This results in a rather complex mosaic of averaged median performances. Still, distinct patterns with respect to individual fingerprints or fragmentation rules can be derived: Common to all three fragmentation rules is a poor performance of the Topological Torsions fingerprint, independent of similarity threshold or scoring function coefficients. A slightly better-than-random performance can only be found for DAIM|torsions|0.1|3.5-0.5::0.51. Especially in case of the RECAP rules – where the best performing sets of parameters were discovered – the performance of Topological Torsions is consistently low (best: RECAP|torsions|0.1|3.5-0.5::0.47). RedFrag's low retrieval rates with the Topological Torsions fingerprint are congruent with short scoring lists, *i.e.*, the number of molecules classified as similar, as depicted in **Figure 3B**. For Topological Torsions, on average, our algorithm leads to 7,995 scorings (RECAP|torsions|0.1) and less than 3,000 scorings for similarity thresholds greater than 0.1, respectively. In contrast to RECAP, the BRICS and DAIM fragmentation rules allow for a higher rate of scoring events for Topological Torsions: 11,195 and 11,893, respectively. In the range of 0.6-0.9 of the similarity thresholds, both fragmentation rules show successful retrieval rates one order of magnitude higher than RECAP (BRICS/DAIM/RECAP|torsions|0.9: 2,294, 3,116 and 216, respectively), but still the lowest values across all fingerprints involved. A further point of note is the highest similarity threshold for which a better-than-random performance is achieved. For all binary fingerprints involved, this similarity-stringency-point is located between DAIM|torsions|0.1 and DAIM/BRICS|topo|0.4. Greater similarity thresholds, *i.e.*, more stringent matching criteria, lead to a drop in performance and the number of successful scoring events. This is not necessarily correlated: Among a low number of scoring events, many actives can be present, thus a high retrieval can be obtained. It becomes clear from that picture that the fuzziness of an RFGraph comparison (and thus the size of the correspondence graph) is limited to a narrow interval of similarity thresholds. When employing stringent similarity criteria, fuzziness is almost eliminated and the correspondence graphs become sparse or non-existent, hence the shortness of the corresponding scoring lists. Of note, the DAIM fingerprint achieves successful scoring events at the highest similarity thresholds (RECAP/BRICS/DAIM|daim|0.9: 8,032 13,315 12,200, scoring list length, respectively) with a better-than-random performance at the maximum similarity threshold of BRICS|daim|0.8. Seemingly, its non-binary, continuous-valued nature is responsible for this extended capability of successful scoring events.

Numerically, our algorithm performed best at RECAP|maccs|0.1|0.5-3.5 with an averaged median AUC value (over 17 activity classes and 30 retrieval runs per class) of 0.59 (**Figure 4**, red data). In the parametrical vicinity of this best performance, there are other well-performing parameter combinations. All of them are located at the bottom left corner of the data block, corresponding to a high value of F_C and low value of G_C, hence an emphasis on the fingerprint similarity term of the scoring function. The same is true for the DAIM fingerprint but not for the other fingerprints, as they show increasing performance when the graph congruency term of the scoring function is emphasized.

Best performances for BRICS and DAIM fragmentation rules are lower than when using RECAP rules, BRICS|topo|0.2|3.5-0.5::0.55 and DAIM|topo|0.2|3.5-0.5::0.57, respectively. Furthermore, high performances, *i.e.*, red squares, appear less often for BRICS and DAIM fragmentation rules. This can be connected to two topological properties of the corresponding RFGraphs as depicted in **Figure S2**. According to the trend lines of median graph size and mean molecular weight of its fragments, RECAP

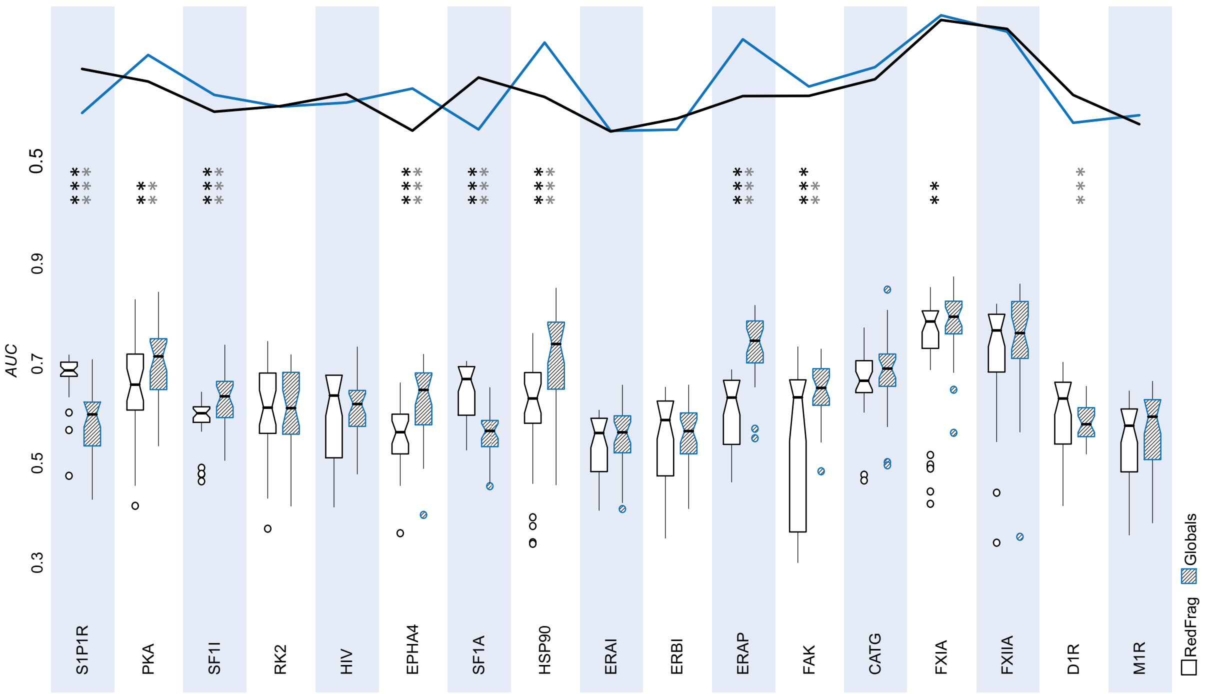


Figure 5: Comparison of RedFrag performance and fingerprint performance without our algorithmic graph reduction (Globals). For both, RedFrag and Globals the top performances for each activity class were selected (OOPS). Black and gray asterisks show the level of statistical significance of the AUC distributions within an activity class as calculated by the t-test and Kolmogorov-Smirnov-test, respectively. ($*** \rightarrow p < 0.01$, $** \rightarrow p < 0.05$, $* \rightarrow p < 0.10$).

fragmentation on average produces smaller RFGraphs consisting of larger chemical epitopes than BRICS and DAIM, the latter working in similar ways. With this fact, we can substantiate the known sensitivity of 2D similarity methods towards higher molecular weights as they correspond to higher property/descriptor richness which in turn allows more discriminative comparisons. Thus, increasing the number of chemical epitopes and simultaneously decreasing their size, *i.e.*, increasing the granularity, reaches a limit where there is no adequate amount of available information for a reasonable similarity calculation. On the contrary, the abandonment of fragmentation, thus regarding a molecule on the whole, is accompanied with the loss of topological and chemical diversity among molecules considered *similar*. In the following, we present and discuss the value of fragmentation as a middle course between atomistic and holistic molecular view.

3.2. Detailed results of the MUV retrieval runs

3.2.1. Impact of fragmentation rules on retrieval performance

With an optimal set of parameters for screening each MUV data set, RedFrag achieves an averaged median AUC value of 0.64 (we will refer to this parameter set as Overall Optimal Parameter Set [OOPS]) (**Figure 5**, black data).

Restricting the choice of parameters to fingerprint, similarity threshold and scoring function coefficients, thus being only fragmentation rule dependent (hereafter Rule Optimal Parameter Set [ROPS]) the performances between the different fragmentation rules were 0.62, 0.59 and 0.60 for RECAP, BRICS and DAIM respectively (**Figure 6**). Inspecting **Figure 6**, there are notable differences between the retrieval rates for each fragmentation pattern: RECAP performs significantly better on the S1P1 receptor, SF1 receptor (SF1I) and HSP 90 ligands, whereas BRICS has a higher retrieval rate on HIV RT-RNase ligands and DAIM on SF1 receptor agonists (SF1A). All three fragmentation patterns have their peak performances with the

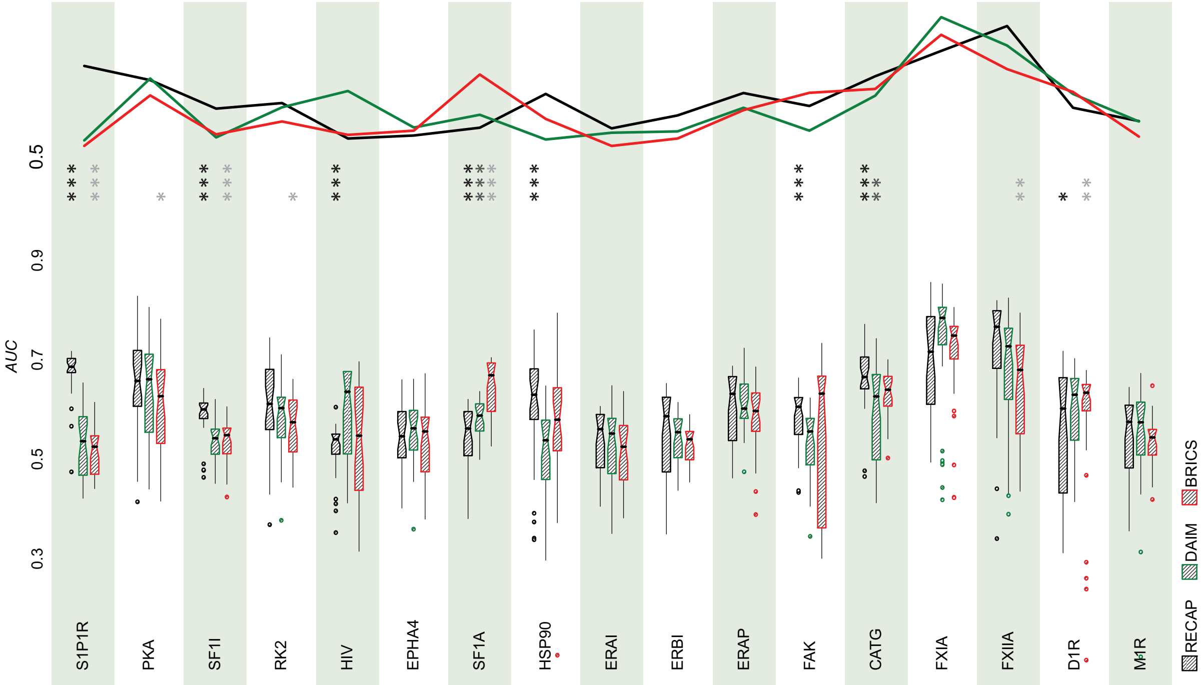


Figure 6: Comparison of RedFrag performance and fingerprint performance without our algorithmic graph reduction (Globals). For both, RedFrag and Globals the top performances for each activity class were selected (OOPS). Black and gray asterisks show the level of statistical significance of the AUC distributions within an activity class as calculated by the t-test and Kolmogorov-Smirnov-test, respectively. (∗∗∗→ $p < 0.01$, ∗∗→ $p < 0.05$, ∗→ $p < 0.10$).

ligands/decoy sets of the coagulation factors FXIa and FXIIa (FXIa::0.71/0.75/0.78 and FXIIa::0.76/0.68/0.72 for RECAP, BRICS and DAIM, respectively). Lowest retrieval rates can be observed for Eph receptor A4, ER-α- and ER-β-coactivator binding inhibitors (EPHA4::0.54/0.56/0.55, ERAI::0.56/0.52/0.55, ERBI::0.58/0.54/0.55). In summary, at the level of single activity classes, no connection between RedFrag's performance and the properties of RFGraphs as depicted in **Figure S2** can be deduced. The outcome of a retrieval experiment is apparently highly dependent on the composition of an activity class with respect to actives and decoys and the fragmentation rule in use.

3.2.2. Impact of fingerprint and scoring-term weighting on the retrieval performance

RedFrag operates consistently well over a broad range of similarity thresholds, scoring function coefficients and fingerprints, as shown by the areas densely occupied by reddish squares in **Figure 3A**. The best performance for a single set of parameters can be observed at RECAP|maccs|0.1|0.5-3.5::0.59 (red data, **Figure 4**). If we look for peak performances for each class across all fragmentation rules, fingerprints, similarity cutoffs and scoring function coefficients, the overall performance increased to 0.64 (black data in **Figure 4**). Especially the S1P1 receptor, SF1 receptor inhibitors and agonists (SF1I and SF1A), HIV RT-RNase and – to a minor extent – ER-α-coactivator binding inhibitors (ERAI) were positively influenced by adapting parameters. As described before, most peak performances are obtained with RECAP fragmentation rules and can be found within the maccs data block. Among the significantly better performances, a transition from highly emphasized fingerprint term to highly emphasized graph congruency term of the scoring function can be observed. A focus on the topological graph similarity is apparently advantageous for these activity classes.

3.2.3. The value of fragmentation

With the MUV data sets, we included not only the idea to evaluate the retrieval effectiveness of actives from a mixed actives-decoys data set, but also wished to investigate whether our algorithm could compete with highly optimized commonly used 2D fingerprints. We compared retrieval performances with and without graph reduction,

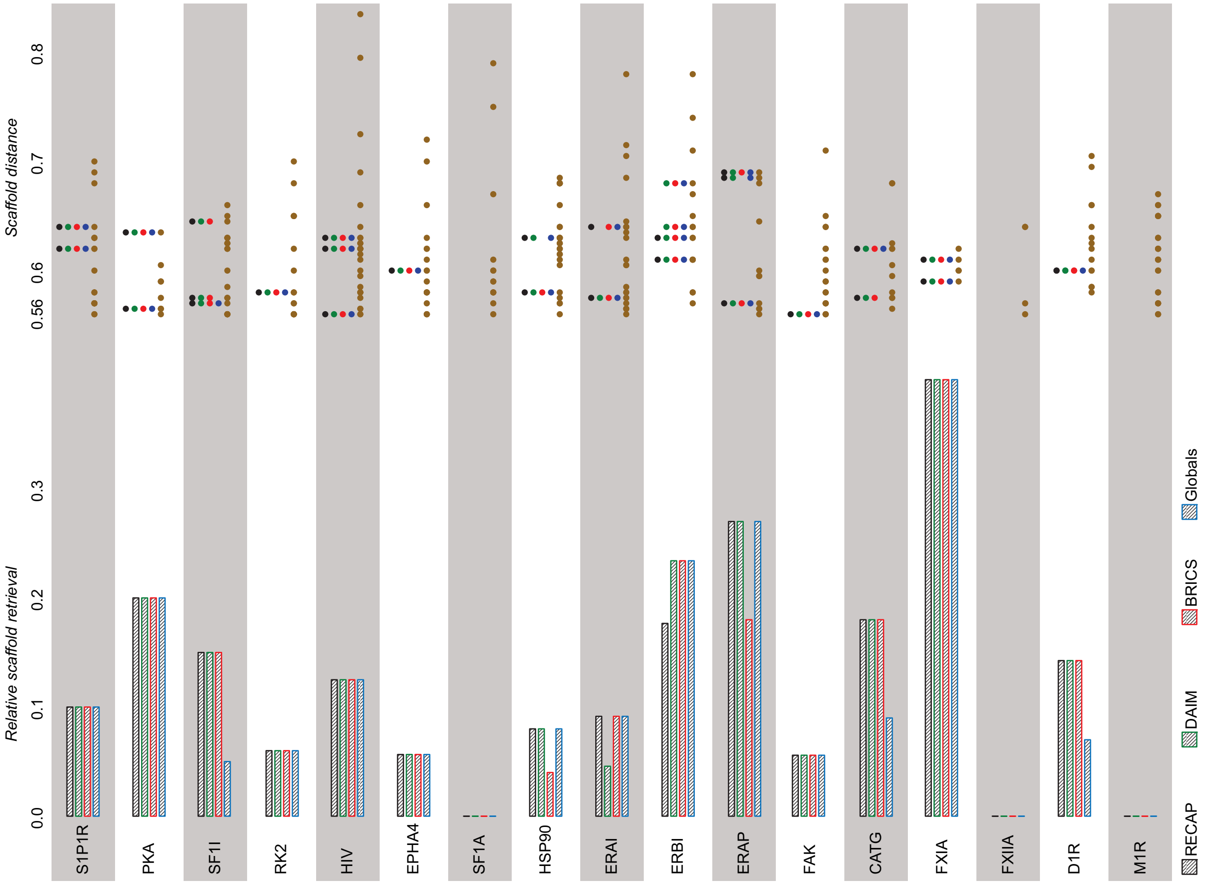
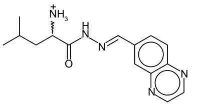
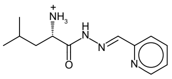
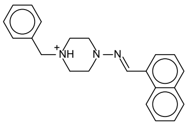
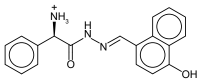
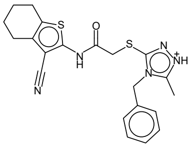
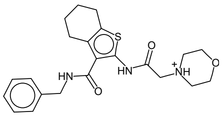
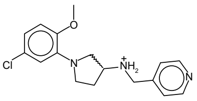
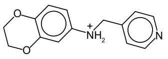
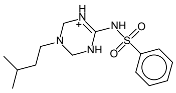
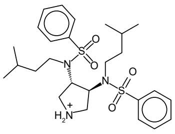
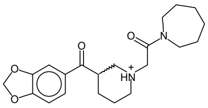
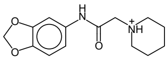


Figure 7: Each activity class was analyzed according to the scaffold distance distribution of the active compounds. Golden circles represent active compounds with an average scaffold distance to any other active greater than the *scaffold distance threshold* (0.56). Colored bars show the relative retrieval performance of *golden scaffolds* for RedFrag and without our algorithmic graph reduction (Globals) both at OOPS.

thus determining the value of fragmentation. In order to make for a comparable and fair approach, for each fingerprint we performed the same validation cycle as in case of RedFrag: Over 17 activity classes with 30 retrieval runs per each class, an averaged median AUC value was calculated for each fingerprint listed in **Table 1**. The results of the unaltered fingerprints (**Figure 3A**), denoted as Globals, show a moderate to good performance for Atom Pairs and Topological Torsions fingerprints, the latter being surprising, as it performed worst when employed within our algorithm. However, we speculate that the poor performance of RedFrag with the Topological Torsions fingerprint can be partially understood given the fact that fragmentation occurs at rotatable bonds, thus effectively reducing the amount of descriptive information defined by successively connected atom quadruples. Hence, the information density of the Topological Torsions fingerprints of the corresponding fragments is reduced, *i.e.*, lowering its discriminative capabilities. In contrast, for MACCS Keys, substructural motifs remain mostly intact even after fragmentation allowing reasonable similarity calculations. A close look at **Figure S3** – where the AUC distributions of all seven fingerprints for each activity class are plotted – shows similar complex performance fluctuations dependent on fingerprint and activity class. As already indicated by the numerical values, Atom Pairs and Topological Torsions fingerprints show the highest retrieval effectiveness for most activity classes, Atom Pairs being at a peak in 8 and Topological Torsions in 5 out of 17 cases. **Figure 5** shows the comparison of peak performances of RedFrag and fingerprints without algorithmic graph reduction (both methodologies at OOPS). From the trends of the median AUC values, a significantly better performance of RedFrag can be seen for the S1P1 receptor, SF1 receptor agonists and D1 receptor ligands. In 7 out of 17 activity classes, RedFrag

Table 3: IC_{50} values, calculated K_i values and ligand efficiencies LE for six experimentally validated hits identified as inhibitors of endothiapepsin.

ID	Hit structure	Query structure	IC_{50} [μM]	K_i [μM][b]	LE[c]
1			4.42 ±0.14	2.08 ±0.07	0.35
2			200.4 ±13.77	94.31 ±6.48	0.22
3			257.2 ±43.87	121.0 ±20.64	0.18
4			301.9 ±81.80	142.1 ±38.49	0.23
5			340.9 ±452.2[a]	160.4±212.8[a]	0.25[a]
6			354.8 ±100.0	166.9 ±47.08	0.18

[a] IC_{50} measurements did not reach plateau. Thus, corresponding values are unreliable and the compound is considered to be a weak binder.
[b] K_i values were calculated using the Cheng-Prusoff equation. (52)
[c] LE is the ligand efficiency, where N is the heavy atom count. [LE] = kcal mol^{-1} N^{-1}

performed equally well as the unmodified fingerprints. In only two cases, HSP 90 and ER-α-coactivator binding potentiators, raw fingerprints outperformed RedFrag. Of note, RedFrag performs better than the MOE molecular properties descriptor (*Molecular Operating Environment, 2007.09*; Chemical Computing Group: Montreal, Canada, 2007) (MOE::0.54) and the SESP descriptor (53) (SESP::0.51) in terms of overall retrieval performance. Both descriptors were used in a virtual screening in order to assess the utility of the compiled MUV data sets in the original publication. (31)

3.2.4. Diversity of enriched compounds

Figure 7 shows the relative retrieval rate of distant scaffolds for each activity class at OOPS compared with retrievals done without algorithmic graph reduction, at OOPS as well. Similar retrieval rates were achieved when using RECAP, BRICS and DAIM decomposition rules as well as experiments without decomposition (denoted as Globals). Moreover, in all cases identical distant scaffolds were retrieved using different fragmentation rules and fingerprints. Of note, however, in the case of SF1 receptor inhibitors (SF1I), Cathepsin G (CATG) and D1 receptor (D1R), respectively, RedFrag retrieved two or one more distant scaffold than fingerprint-only enrichment. As for the retrieval performance in general, RECAP fragmentation presents itself as an advantageous choice for the enrichment of structurally diverse actives.

3.3. Prospective study

A pre-inspection of the cluster representatives followed the sphere exclusion clustering of the initial virtual screen. This resulted in a set of 48 compounds. The number of cluster representatives was reduced by removing molecules that were considered trivial matches or were based on incorrect input structures. Further selection based on diversity considerations reduced this set to 20 compounds which were purchased and tested for their activity towards endothiapepsin.

3.3.1. The endothiapepsin system

Endothiapepsin is involved in a wide range of diseases such as hypertension and malaria. (54) It belongs to the family of pepsin-like aspartic proteases and has been used as a model enzyme for not only mechanistic studies (55,56,57), but also for the development of renin (58) and β-secretase (59) inhibitors.

3.3.2. Experimental validation of virtual hits

Three out of twenty predicted compounds inhibit endothiapepsin with IC_{50} values in the range of 4.42 to 257 µM. The most potent inhibitor, **1**, a close analog to one of our query molecules, **Q4** (**Table S42**), features an IC_{50} value of 4.42 µM. Three more compounds showed inhibition in the three-digit µM-range, one of them without reaching a plateau, and thus without a reliable IC_{50} value, **Table 3**.

3.3.3. Diversity and scaffold hopping analysis

Figure S4 shows the similarity matrix based on the ECFP4 fingerprint between our 20 selected virtual hits and the query molecules employed for the virtual screening. Among the six experimentally validated hits, only one compound (**1**) shows a high ECFP4 Tanimoto similarity to two of the queries. Our hit shares a leucine-hydrazone-moiety with the queries and differs by a larger quinazoline moiety, whereas the query molecules contain a trifluorotoluene- and pyridyne moiety at this position, respectively. Interestingly, a closely related virtual hit (**17**) with a methoxyquinoline moiety did not show any inhibition towards endothiapepsin. **Figure S5** shows the scaffold similarity matrix between the virtual and validated hits and the query molecules. Compound **2** shows moderate activity and high ECFP4 and scaffold similarity to only one query molecule (**Q24**) which is due to the naphthyl moiety. Compound **6**, which is only a weak binder to endothiapepsin, is a dissimilar compound measured by both ECFP4 and scaffold similarity. The most active compound along with its inactive but close analog (**1** and **17**) reveal a close distance to queries **Q4** and **Q5** by means of ECFP4 and, interestingly, to **Q11** based on scaffold similarity. According to the latter, compound **1** can be considered a scaffold hop, due to mediocre graph similarity and high chemical diversity. Compound **3**, which is again a moderate binder, shows neither ECFP4 nor scaffold distance relationship to any of the query molecules, leading to a true scaffold hop, as is true for compound **6**.

4. Discussion and conclusions

RedFrag is based in the premise that fragment-based searches with infinite colors are a fast, flexible and intuitive similarity measure that is also capable of recognizing chemically equivalent, yet topologically different fragments in two molecules. We have tested this method retrospectively on the MUV data sets, to avoid distortion of the performance due to trivial separations of actives and decoys. Going beyond retrospective analysis, we also applied our method to predict binders for the aspartic protease endothiapepsin. RedFrag performed at the same level as the other methods investigated on the MUV data set. In some cases, existing algorithms were surpassed while a certain level of underperformance has to be noted in a few others. Strikingly, and within rather broad limits, the performance of RedFrag does not depend on the parameters chosen. We can thus conclude that the parameters selected for the final screen likely did not suffer from training set bias, as similar parameters would have been obtained going out from a different molecule set. **Figure 3A** and **B** allow us to draw another interesting conclusion: despite the simplicity of the DAIM fingerprint, the performance of RedFrag does not deteriorate substantially, which again argues for the stability of the general approach. One can also speculate why the RECAP fragmentation pattern performs best: It generates the largest fragments of the three fragmentation patterns. Especially in cases where the molecules contain many potential fragments, this could help by generating feature-rich fragments, which then means that matches involving such fragments have a relatively larger weight. Nevertheless, BRICS and DAIM fragmentation patterns are justified as they partially cover an orthogonal set of disconnection sites, allowing for the decomposition of chemically diverse and optimized compounds in the field of medicinal chemistry. Particularly for DAIM decomposition rules, where rotatable bonds are cut independent of retrosynthetic considerations, deviating topology or single bioisosteric replacements of functional groups should not be an insurmountable challenge. RedFrag also has the potential to retrieve scaffolds distant from the query, *i.e.* perform scaffold hops. This can be gleaned from **Figure 7**, where the propensity of RedFrag to retrieve unusual scaffolds is comparable to other methods. This ability of RedFrag is also nicely demonstrated in the prospective screen, where the second and third most potent compounds are distant from the known ligand space of endothiapepsin, as evaluated by ECFP4 fingerprints. At the same time, the prospective screen also yielded a very potent molecule, which at 4 µM is one of the more potent binders of endothiapepsin. While it cannot be described as novel when compared to the molecules that were used as query, it is a clear improvement on the compound series published earlier (Mondal *et al.* (45)). Further comparison between the similarity matrices in **Figure S4**, **Figure S5** and **Figure S6** shows that the answer to what is similar given by the RedFrag approach is different from that given based on the ECFP4 fingerprint or derived from scaffold frameworks. Thus, it seems likely to find a higher overall number of actives by applying different search algorithms.

Of note, RedFrag is also fast: The screen of the 7.3 million molecules in the ZINC library took less than 8 hours on a single core CPU. Such a speed is of the same order as other fingerprint-based methods, but an order of magnitude faster than currently possible with docking calculations.

In conclusion, we demonstrated that RedFrag is an intuitive, fast and unbiased algorithm for 2D molecule searches while it only has two main parameters and needs little optimization. The performance in both retrospective and prospective studies is in line with existing methods, but often yields different answers, adding it to the repertoire of suitable approaches for large-scale screenings. Most importantly, we abolished the need for pre-defining a color space without a loss in performance. This means that our algorithm is applicable to all sets of molecules without restrictions. At the same time, RedFrag is “plug and play” in the sense that future decomposition rules or fingerprints can be effortlessly integrated in the basic approach.

5. Acknowledgments

Peter Kolb thanks the German Research Foundation DFG for Emmy Noether fellowship K04095/1-1.

References

1. Irwin, J. J.; Shoichet, B. K. ZINC -- A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005,** *45,* 177-182.

2. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012,** *40* (D1), D1100-D1107.

3. Lajiness, M. S. Dissimilarity-based compound selection techniques. *Perspectives in Drug Discovery and Design* **1997,** *7* (1), 65-84.

4. Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002,** *45,* 4350-4358.

5. Willett, P.; Wild, D. J. Similarity Searching in Files of Three-Dimensional Chemical Structures. Alignment of Molecular Electrostatic Potential Fields with a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1996,** No. 36, 159-167.

6. Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998,** No. 38, 983-996.

7. Hubalek, Z. Coefficients of Association and Similarity, Based on Binary (Presence-Absence) Data: an Evaluation. *Biol. Re. Cambridge Philos. Soc.* **1982,** No. 57, 669-689.

8. Gower, J. C. Measures of Similarity, Dissimilarity and Distance. Wiley: Chester, U.K., 1982.

9. Hagadone, T. R. Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases. *J. Chem. Inf. Comput. Sci* **1992,** *32* (5), 515-521.

10. Guner, O. F. History and Evolution of the Pharmacophore Concept in Computer-Aided Drug Design. *Current Topics in Medicinal Chemistry* **2002,** *2* (12), 1321-1332.

11. Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *Journal of Medicinal Chemistry* **2007,** *50,* 74-82.

12. Cramer, R. D.; Jilek, R. J.; Andrews, K. M. Dbtop: topomer similarity searching of conventional structure databases. *J. Mol. Graph. Model.* **2002,** *20* (6), 447-462.

13. Cramer, R. D.; Jilek, R. J.; Gaussregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. "Lead hopping". Validation of topomer similarity as a superior predictor of similar biological activities. *J. Med. Chem.* **2004,** *47* (27), 6777-6791.

14. Schuffenhauer, A. Computational methods for scaffold hopping. *WIREs Comput. Mol. Sci.* **2012,** No. 2, 842-867.

15. Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985,** *25* (2), 64-73.

16. Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987,** *27* (2), 82-85.

17. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010,** *50* (5), 742-754.

18. Gobbi, A.; Poppinger, D. Genetic optimization of combinatorial libraries. *Biotechnol. Bioeng.* **1998,** *61* (1), 47-54.

19. Kolb, P.; Caflisch, A. Automatic and efficient decomposition of two-dimensional structures of small molecules for fragment-

based high-throughput docking. *J. Med. Chem.* **2006,** *49,* 7384-7392.

20. Rarey, M.; Dixon, J. S. Feature trees: A new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Design* **1998,** *12* (5), 471-490.

21. Rarey, M.; Stahl, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001,** No. 44, 1035-1042.

22. Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003,** *43* (2), 338-345.

23. Takahashi, Y.; Sukekawa, M. Automatic identification of molecular similarity using reduced-graph representation of chemical structure. *J. Chem. Inf. Comput. Sci.* **1992,** *32* (6), 639-643.

24. Barker, E. J.; Gardiner, E. J.; Gillet, V. J.; Kitts, P.; Morris, J. M. Further Development of Reduced Graphs for Identifying Bioactive Compounds. *J. Chem. Inf. Comput. Sci.* **2003,** *43* (2), 346-356.

25. Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999,** *19,* 2894– 2896.

26. Brown, N.; Jacoby, E. On Scaffolds and Hopping in Medicinal Chemistry. *Mini-Rev. Med. Chem.* **2006,** *6,* 1217– 1229.

27. Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump? *QSAR Comb. Sci.* **2006,** *25,* 1162– 1171.

28. Hu, Y.; Bajorath, J. Global Assessment of Scaffold Hopping Potential for Current Pharmaceutical Target. *Med. Chem. Commun.* **2010,** *1,* 339– 344.

29. Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010,** *50,* 205– 216.

30. Stumpfe, D.; Bajorath, J. Applied Virtual Screening: Strategies, Recommendations, and Caveats. In *Methods and Principles in Medicinal Chemistry. Virtual Screening. Principles, Challenges, and Practical Guidelines;* Sotriffer, C., Ed.; Wiley-VCH: Weinheim, 2011; pp 73– 103.

31. Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioact ivity Data. *J. Chem. Inf. Model.* **2009,** *49* (2), 169-184.

32. Li, R.; Stumpfe, D.; Vogt, M.; Geppert, H.; Bajorath, J. Development of a Method To Consistently Quantify the Structural Distance between Scaffolds and To Assess Scaffold Hopping Potential. *J. Chem. Inf. Model.* **2011,** *51* (10), 2507-2514.

33. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAPRetrosynthetic Combinatorial Analysis Procedure:â€‰ A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998,** *38* (3), 511-522.

34. Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *Current Medicinal Chemistry* **2008,** *3* (10), 1503-1507.

35. RDKit. Open source toolkit for chemoinformatics. **2000-2013**.

36. NetworkX. Open source package for creation, manipulation, and study of the structure, dynamics, and functions of complex networks. **2004-2012**.

37. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding

rules. *J. Chem. Inf. Comput. Sci.* **1988,** *28* (1), 31-36.

38. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006,** *11,* 1046-1053.

39. Levi, G. A note on the derivation of maximal common subgraphs of two directed. *Calcolo* **1972,** *9,* 341-352.

40. Borrow, H.; Burstall, R. Subgraph isomorphism, matching relational. *Inf. Proc. Lett.* **1976,** *4,* 83-84.

41. Raymond, J. W.; Gardiner, E. J.; Willett, P. RASCAL: Calculation of graph similarity using maximum common edge subgraphs. *The Computer Journal* **2002,** *45,* 631-644.

42. Stahl, M.; Mauser, H.; Tsui, M.; Taylor, N. R. A Robust Clustering Method for Chemical Structures. *J. Med. Chem.* **2005,** *48* (13), 4358–4366.

43. Irwin, J. J.; Sterlin, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J Chem Inf Model* **2012,** *52* (7), 1757-1768.

44. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucl. Acids Res.* **2012,** *40* (D1), D1100-D1107.

45. Mondal, M.; Radeva, N.; Köster, H.; Park, A.; Potamitis, C.; Zervou, M.; Klebe, G.; Hirsch, A. K. H. Structure-Based Design of Inhibitors of the Aspartic Protease Endothiapepsin by Exploiting Dynamic Combinatorial Chemistry. *Angewandte Chemie International Edition* **2014,** *53* (12), 3259-3263.

46. Köster, H. *Endothiapepsin und Proteinkinase A: Komplexstrukturen mit neuartigen Inhibitoren, Durchmustern einer Fragmentbibliothek sowie Inhibitordesign ausgehend von einer Sonde;* Philipps-Universität Marburg: Marburg, 2012.

47. Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999,** *39* (4), 747-750.

48. Toth, M. V.; Marshall, G. R. A simple, continuous fluorometric assay for HIV protease. *International Journal of Peptide and Protein Research* **1990,** *36* (6), 544-550.

49. Köster, H.; Craan, T.; Brass, S.; Herhaus, C.; Zentgraf, M.; Neumann, L.; Heine, A.; Klebe, G. *J. Med. Chem.* **2011,** *54,* 7784-7796.

50. JChem 15.4.27.0 2015 ChemAxon. **2015**.

51. OpenEye Toolkits 2015.Feb.3. **2015**.

52. Cheng, H. C. The power issue: determination of KB or Ki from IC50: A closer look at the Cheng–Prusoff equation, the Schild plot and related power equations. *Journal of Pharmacological and Toxicological Methods* **2001,** *46* (2), 61-71.

53. Baumann, K. An Alignment-Independent Versatile Structure Descriptor for QSAR nad QSPR Based on the Distribution of Molecular Features. *J. Chem. Inf. Comput. Sci.* **2002,** *42,* 26-35.

54. Cooper, J. B. Aspartic Proteinases in Disease: A Structural Perspective. *Current Drug Targets* **2002,** *3* (2), 155-173.

55. Coates, L.; Erskine, P. T.; Wood, S. P.; Myles, D. A. A.; Cooper, J. B. A Neutron Laue Diffraction Study of Endothiapepsin: Implications for the Aspartic Proteinase Mechanism. *Biochemistry* **2001,** *40* (44), 13149-13157.

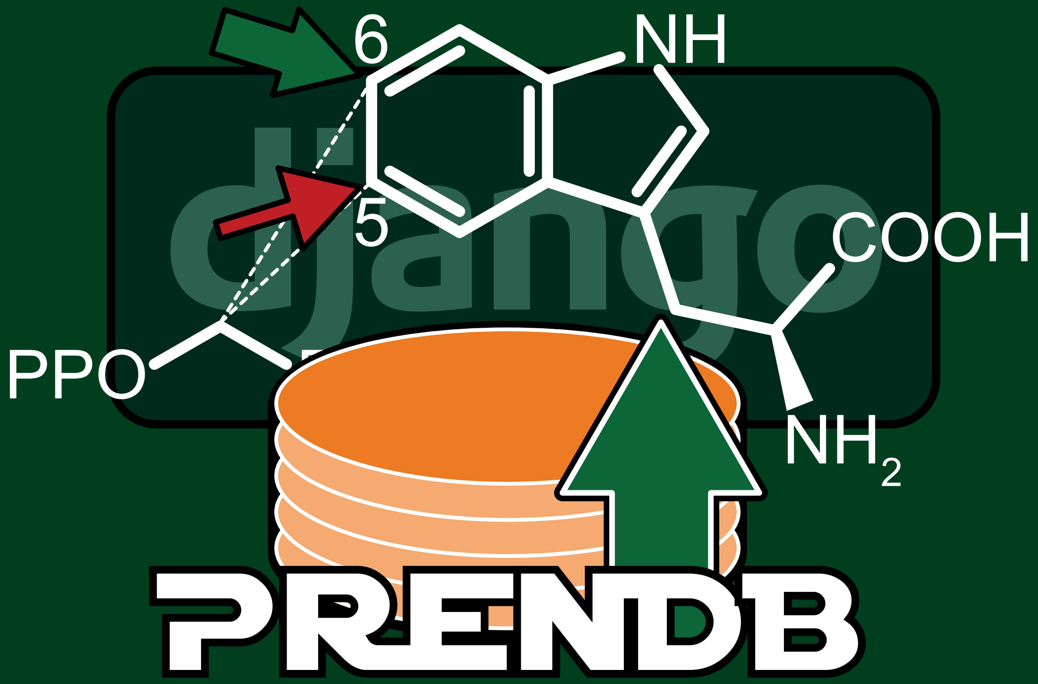
56. Coates, L.; Erskine, P. r.; Mall, S.; Gill, R.; Wood, S. P.; Myles, D. A. A.; Cooper, J. B. X-ray, neutron and NMR studies of the catalytic mechanism of aspartic proteinases.

European Biophysics Journal **2006,** *35* (7), 559-566.

57. Coates, L.; Tuan, H.-F.; Tomanicek, S.; Kovalevsky; rey; Mustyakimov, M.; Erskine, P.; Cooper, J. The Catalytic Mechanism of an Aspartic Proteinase Explored with Neutron and X-ray Diffraction. *J. Am. Chem. Soc.* **2008,** *130* (23), 7235-7237.

58. Cooper, J.; Quail, W.; Frazao, C.; Foundling, S. I.; Blundell, T. L.; Humblet, C.; Lunney, E. A.; Lowther, W. T.; Dunn, B. M. X-ray crystallographic analysis of inhibition of endothiapepsin by cyclohexyl renin inhibitors. *Biochemistry* **1992,** *31* (35), 8142-8150.

59. Geschwindner, S.; Olsson, L.-L.; Albert, J. S.; Deinum, J.; Edwards, P. D.; de Beer, T.; Folmer, R. H. A. Article Previous Article Next Article Table of Contents Discovery of a Novel Warhead against β-Secretase through Fragment-Based Lead Generation. *J. Med. Chem.* **2007,** *50* (24), 5903-5911.



Journal of Biological Chemistry **2017**, *292*, 4003-4021.

Part 2

Part 2 encloses the full article which is intended to be the prelude to our *SAR-by-Enzyme* study, where we exploit the biochemically and medicinal-chemically interesting catalytic property of dimethylallyltryptophan synthases: An enzyme family which throughout species plays a key role in the diversification of secondary metabolites. In order to use these enzymes as biocatalytic tools for chemical transformation of lead- or drug-like chemical matter we firstly collected reaction data from public domain and cast it into a database of prenyltransferase reactions. This database not only contains detailed information about the reaction of a specific prenyltransferase but also it enables the user to predict putative substrates for a prenylation reaction.

The author list is the following (by contribution order): Gunera J., Kindinger F., Li S.M. and Kolb P. I was responsible for the design and layout of the database, the automated analysis of reaction data extracted from literature, its integration into the database, programming of the predictive subroutines therein and the hierarchical virtual screen of putative substrates. Florian, who equally contributed to this study, designed and performed the validation enzyme assays, analyzed the prenylation results, extracted reaction products and elucidated their structure.

PrenDB: A substrate prediction database to enable biocatalytic use of prenyltransferases

Jakub Gunera[‡,†,1], Florian Kindinger[§,1], Shu-Ming Li[§,†] and Peter Kolb[‡,†]

[‡]Department of Pharmaceutical Chemistry, Philipps-University, Marburg, Hesse, 35032, Germany

[§]Department of Pharmaceutical Biology and Biotechnology, Philipps-University, Marburg, Hesse, 35032, Germany

[†]Synmikro, LOEWE Centre for Synthetic Microbiology, Philipps-University, Marburg, Hesse, 35043, Germany

[1]These authors contributed equally to this study.

Keywords:

DMATS prenyltransferases, database, repitope, subgraph isomorphism

Abstract

Prenyltransferases of the ***d****i****m****ethyl****a****llyl****t****ryptophan* ***s****ynthase (DMATS) superfamily catalyze the attachment of prenyl or prenyl-like moieties to diverse acceptor compounds. These acceptor molecules are generally aromatic in nature and mostly indole or indole-like. Their catalytical transformation represents a major skeletal diversification step in the biosynthesis of secondary metabolites including the indole alkaloids. DMATS enzymes thus contribute significantly to the biological and pharmacological diversity of small molecule metabolites. Understanding the substrate specificity of these enzymes could create opportunities for their biocatalytic use in preparing complex synthetic scaffolds. However, there has been no framework to achieve this in a rational way. Here, we report a chemoinformatic pipeline to enable prenyltransferase substrate prediction. We systematically cataloged 32 unique prenyltransferases and 167 unique substrates to create possible reaction matrices, and compiled these data into a browsable database named PrenDB. We then used a newly developed algorithm based on molecular fragmentation to automatically extract reactive chemical epitopes. The analysis of the collected data sheds light on the thus far explored substrate space of DMATS enzymes. To assess the predictive performance of our virtual reaction extraction tool, 38 potential substrates were tested as prenyl acceptors in assays with three prenyltransferases, and we were able to detect turnover in more than 61 % of the cases. The database, PrenDB (www.kolblab.org/prendb.php), enables the prediction of potential substrates for chemoenzymatic synthesis through substructure similarity and virtual chemical transformation techniques. It aims at making prenyltransferases and their highly regio- and stereoselective reactions accessible to the research community for integration in synthetic workflows.*

1. Introduction

Prenylated primary and secondary metabolites including indole alkaloids, flavonoids, coumarins, xanthones, quinones and naphthalenes are widely distributed in terrestrial and marine organisms. They exhibit a wide range of biological activities including cytotoxic, antioxidant, and antimicrobial activities. (1,2,3) Compared to their non-prenylated precursors,

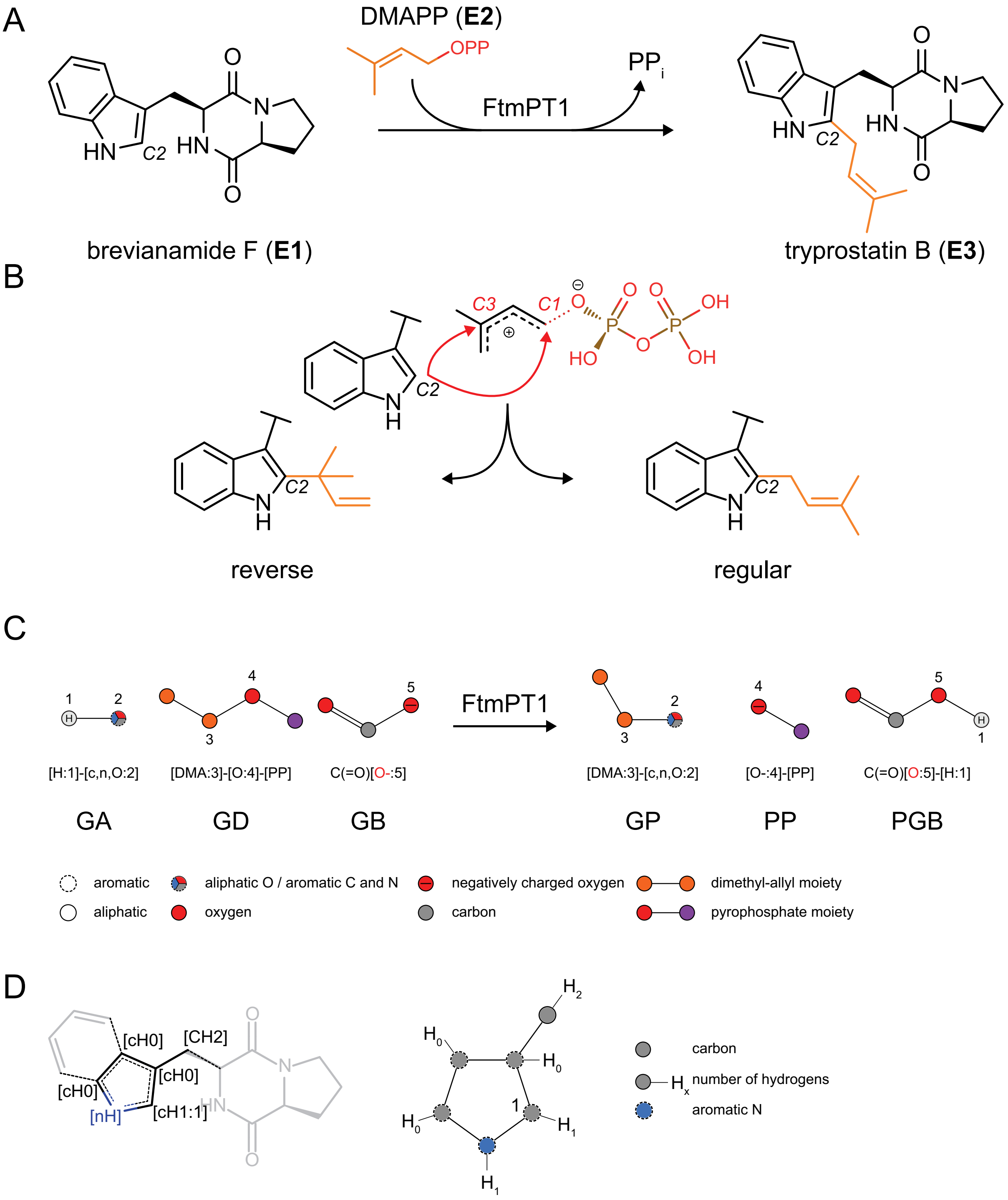


Figure 1: **A** Exemplary transformation of brevianamide F (**E1**) to tryprostatin B (**E3**). **B** Regiochemistry of the nucleophilic attack on the prenyl moiety (**E2**). For regular prenylation, bond formation occurs between *C2* and the carbon adjacent to the pyrophosphate group. An attack of *C2* on the tertiary carbon of DMAPP leads to the reversely prenylated product. **C** Illustration of the SMIRKS-like notation derived from **A** (generated by SMARTSviewer) (23). **GA** = general prenyl moiety acceptor; **GD** = general prenyl moiety donor; **GB** = general base; **GP** = general prenylation product; **PP** = pyrophosphate; **PGB** = protonated general base. **D** (left) A reactive epitope indicated around the reactive atom ([cH1:1]) with the atomic properties given in SMARTS nomenclature (brackets). **D** (right) Reactive epitope as generated by SMARTSviewer. (23)

these compounds usually demonstrate distinct and often improved biological and pharmacological activities, which makes them promising candidates for drug discovery and development. (1,2,4,5,6) These compounds could be considered hybrid molecules of prenyl moieties of different chain lengths (n·C5, where n is an integer number) and aromatic skeletons originating from various biosynthetic pathways. (7,8) Prenyl transfer reactions, *i.e.*, the connections of prenyl moieties to the aromatic nucleus, are catalyzed by a diverse family of prenyltransferases. Interestingly, this step usually represents the key transformation in the biosynthesis of such compounds. A prenyl moiety can be attached by prenyltransferases via its *C1* (regular prenylation) or *C3* (reverse prenylation) to *C*-, *O*- or *N*-atoms of an acceptor (**Figure 1A** and **B**). (7,8) Together with the observed regiospecific prenylations at different positions of an acceptor molecule, prenyltransferases contribute significantly to the structural and biological diversity of natural products. (7)

Based on their amino acid sequences, biochemical and structural characteristics, prenyltransferases are categorized into different subgroups. (7) In the last decade, significant progress has been achieved with the members of the DMATS (dimethylallyltryptophan synthase) superfamily and more than 40 enzymes of this group were identified and characterized by mining of fungal and bacterial genomes. (7) These enzymes catalyze transfer reactions of a prenyl moiety from prenyl pyrophosphate, *e.g.*, dimethylallyl pyrophosphate (DMAPP), to diverse acceptors such as tryptophan, tyrosine, tryptophan-containing cyclic dipeptides, xanthones, tricyclic or tetracyclic aromatic moieties or even non-aromatic compounds. Among the acceptors, indole derivatives including tryptophan and tryptophan-containing cyclic dipeptides are substrates of most of the DMATS enzymes investigated so far. (7,9)

The DMATS enzymes have already been demonstrated to display high substrate and catalytic promiscuity. They catalyze not only prenylation of their substrates and closely related compounds, but also use structurally quite different compounds as prenyl acceptors. (10) Therefore, these enzymes were successfully used for production of a large number of prenylated

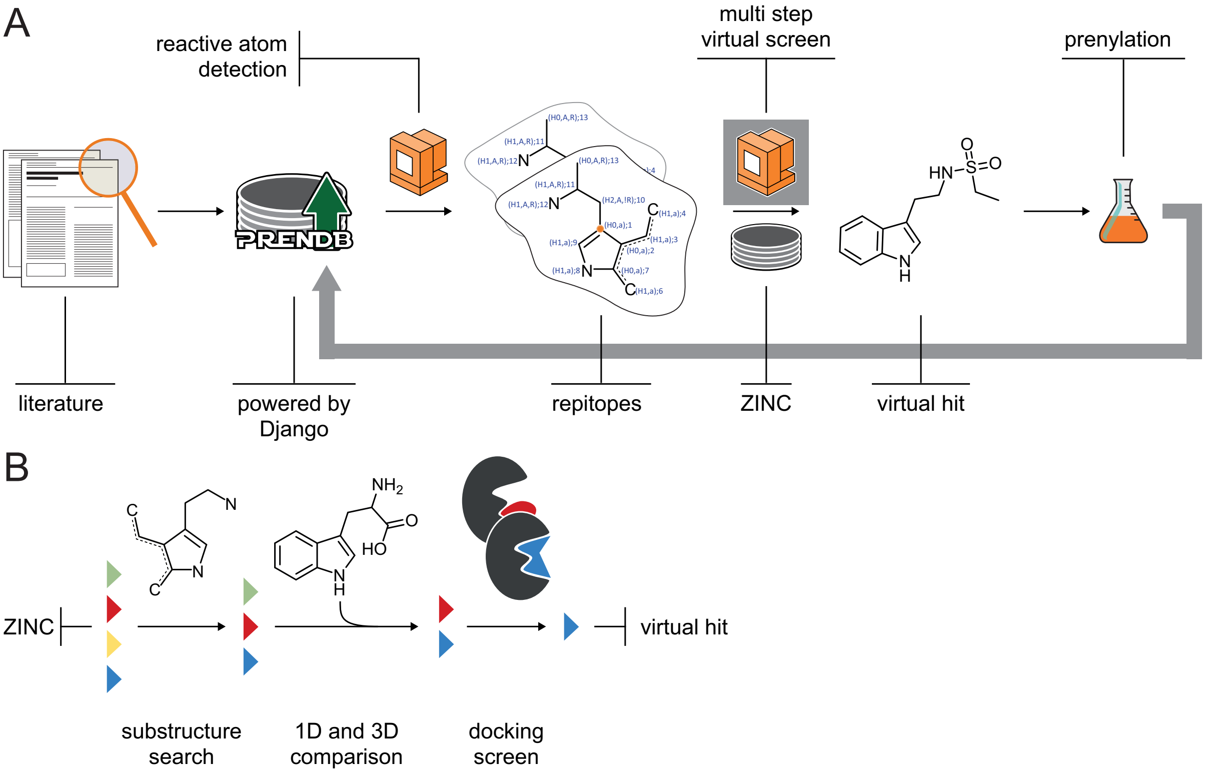


Figure 2: Schematic of the underlying workflow. **A** Reactions of prenyltransferases were digitalized and stored in PrenDB. The reactive atoms were detected algorithmically and reconstituted to reactive epitopes (repitopes). Prenylation candidates were selected by a multi-step virtual screening approach and their transformation potentials were evaluated experimentally on selected prenyltransferases. **B** Multi-step virtual screening of database compounds (colored triangles). i) Repitope-based substructure search based on the substrate space covered in PrenDB; ii) One-dimensional property and three-dimensional shape comparison with known substrates and iv) docking screen on three promiscuous prenyltransferases with available crystal structures.

Figure 3: Illustrative interpretation of the analysis of PrenDB: 32 enzymes contribute to the entirety of digitalized reactions (upper histogram). The reactions can be further subdivided into categories based on the chemistry of the transformation (middle). Each subgroup corresponds to one or more enzymes (lower histogram). *The contribution of the enzyme with the highest contribution to a particular group was set to 100.

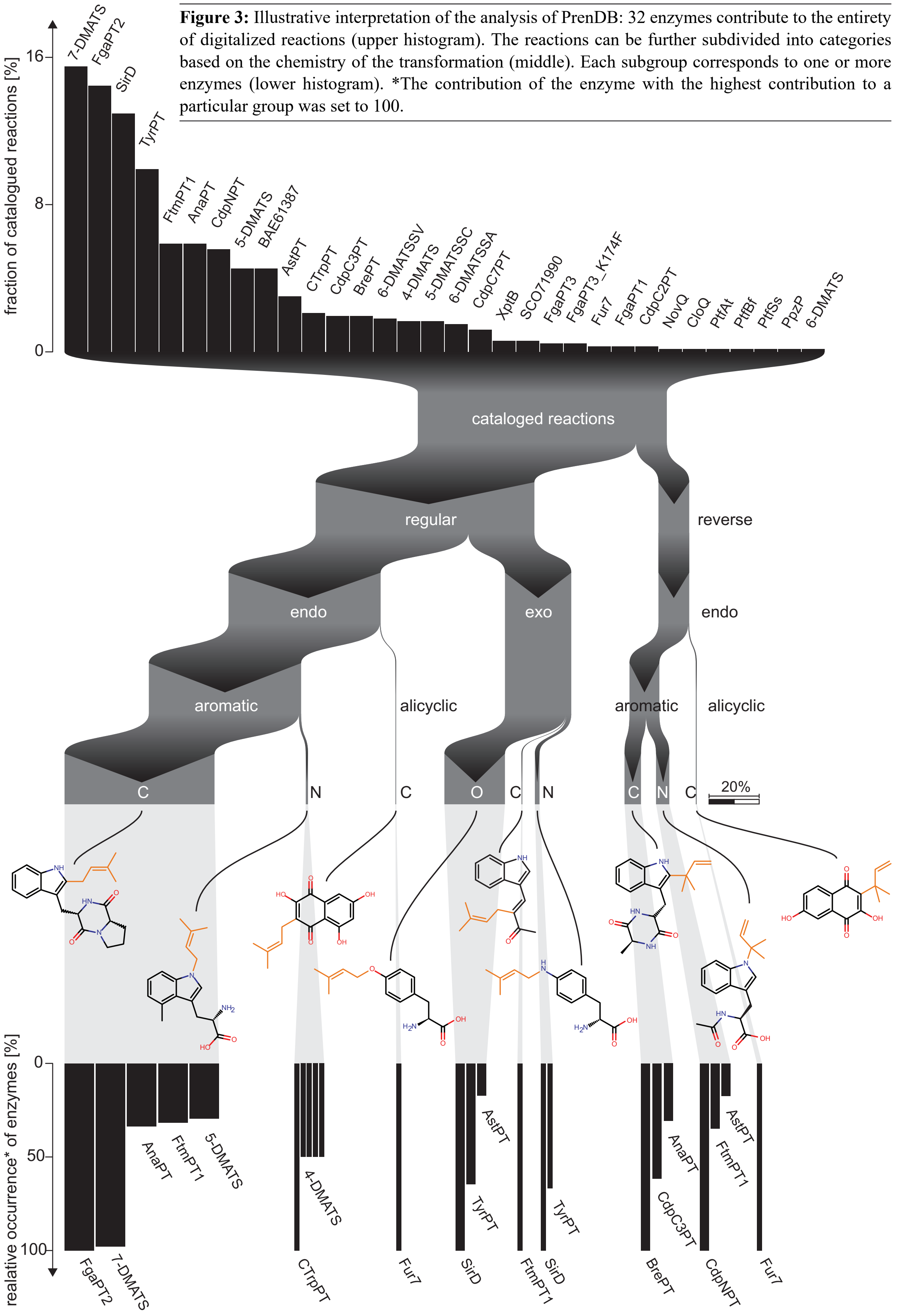
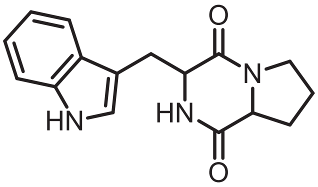
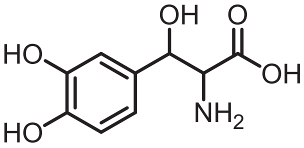
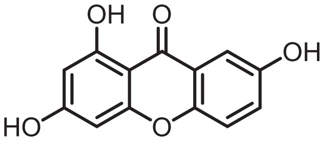
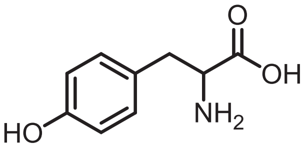
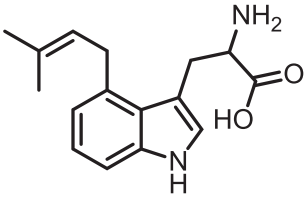


Table 1: Cluster size and substrate space coverage derived from hierarchical clustering.

cluster representative	clusters	size	substrate space coverage [%]	description
	1, 2, 3	38	22	unsubstituted indole moieties and tryptophan-containing cyclic dipeptides
	4, 5	13	8	derivatives of tyrosine: modifications on benzene ring, homologs
	6, 7, 8	45	27	derivatives of xanthone, naphthalene, quinone and flavonoid
	9	18	11	tyrosine derivatives with modifications on benzene ring
	10, 11, 12, 13, 14	53	32	side-chain-modified and multiply prenylated tryptophans

derivatives including prenylated tryptophan and tyrosine analogs, tryptophan-containing cyclic dipeptides and derivatives thereof, hydroxyxanthones, hydroxynaphthalenes, flavonoids, indolocarbazoles and acylphloroglucinols. (10) For example, *N1*-, *C4*-, *C5*-, *C6*- and *C7*-prenylated tryptophan and *N1*-, *C2*-, *C3*-, *C4*- and *C7*-prenylated tryptophan-containing cyclic dipeptides and derivatives were obtained by using DMATS enzymes as biocatalysts. (9,10)

One of the problems to discover and use DMATS enzymes as biocatalysts in a rational and targeted manner is the prediction of the acceptance of a putative substrate. On the one hand, the enzymes share similar structures, albeit often at low sequence identities, and catalyze, in many cases, similar reactions. On the other hand, different enzymes with similar natural substrates accepted further non-native aromatic substances with clearly different activities. (7) Therefore, bioinformatic and chemoinformatic approaches for the prediction of the catalytic activity of these enzymes are welcome and necessary in order to harness the full biosynthetic potential of this enzyme class.

We describe in this work the creation and evaluation of a database that catalogs and stores prenyltransferase reaction information. Because storage of the reactions is automated, the database is not static, but will grow with each new reaction described in the literature. Furthermore, we present an application of PrenDB where we predict and validate putative substrates for prenyltransferases (**Figure 2**).

2. Results

2.1. PrenDB Statistics

Digitalization and chemoinformatic encoding of enzymatic reactions of the DMATS superfamily allows for a deep analysis of their substrate space and reactivity towards distinct chemical epitopes (**Figure 3**). In total, 32 unique enzymes were found throughout the inspected literature. The three most prominent prenyltransferases in terms of the number of annotated transformations are 7-DMATS, FgaPT2 and SirD, accounting for 15, 14 and 13 % of the reactions in the database, respectively. At the other end of the spectrum, there are seven enzymes for which only a single reaction has been published. With respect to promiscuity, the number of unique reactive epitopes – molecular substructures centered around the reactive atom and henceforth called repitopes in this work – was used as a descriptor (*cf.* **Figure 1D** for an exemplary repitope): The enzymes 7-DMATS, FgaPT2 and AnaPT transfer prenyl moieties onto the broadest range of chemical epitopes. Together, these three enzymes contribute more than 65 % to the repitope space. The knowledge of the reactive atom and its surroundings makes a further distinction of enzymatic transformations possible: The Sankey diagram in **Figure 3** shows how the cataloged reactions can not only be linked to their prenyltransferases, but also subdivided into types per the reactive atom. The clear majority of reactions (87 %) corresponds to the regular type of prenyl moiety transfer (**Figure 1B**, right), where the thermodynamically more stable regioisomer is formed.

In 73 % of all reactions and in all reverse attachments, the reactive atom is a member of a ring system (endo). Only a small part of regular prenylations occur at exocyclic atoms (exo, 26 %). There, prenyl moieties are transferred onto oxygen and nitrogen atoms of tyrosine and aniline-like moieties by SirD, FtmPT2 or 4-DMATS. Reverse prenylation can be observed at carbon and nitrogen atoms only. They are incorporated in aromatic ring systems, less frequently also in alicyclic moieties such as benzoquinones. More than 60 % of all reactions occur at aromatic carbon atoms: Derivatives of indole, including tryptophan, are the most frequent repitopes in this largest reaction subclass. Much rarer are prenylations at nitrogen (8 %) or at exocyclic oxygen atoms (23 %). By comparison, 5 % (33 entries) of reactions of tryptophan-like moieties (*e.g.*, at atom position *C2* in brevianamide F (**E1**) (**Figure 1A**) lead to formation of compounds with a fused ring system, where the atomic environment of the reactive atom becomes dearomatized during the reaction (FtmPT1, Fur7, 4-DMATS).

2.1.1. Substrate space

Throughout the analyzed literature – 44 articles from 17 journals – 167 unique substrates were found. In order to analyze the substrate space diversity of prenyltransferases, a similarity matrix based on the pairwise ECFP4 (11) fingerprint molecular similarity was calculated (**Figure 4**, **Table 1** and **Table S1**), followed by a hierarchical clustering. This allows the grouping of substrates based on their chemical structure. From the corresponding dendrograms and supported by the reorganized distance matrix, five substrate classes distributed over 14 clusters can be deduced: i) Unsubstituted indoles, derivatives of tryptophan and proline-tryptophan cyclic dipeptides; ii) Derivatives of tyrosine with modifications on benzene and aliphatic atoms; iii) Naphthalene, quinone and flavonoid derivatives; iv) Side chain modified tyrosines and v) side chain modified and multiply prenylated tryptophans. More than 50 % of the substrate space is covered by indole-containing compounds. Molecules with tyrosine and flavonoid or xanthone motifs contribute 18 % and 26 % to the substrate space, respectively. Furthermore, the space spanned by the fragments obtained via bond cleavage during the fragment-based subgraph isomorphism perception process (*cf.* Materials and Methods) is covered to an extent of 64 % by tryptophan and diketopiperazine epitopes. This predominance of indoles can be explained by tryptophan and indole derivatives being the native substrates for 78 % of the enzymes in PrenDB. This, combined

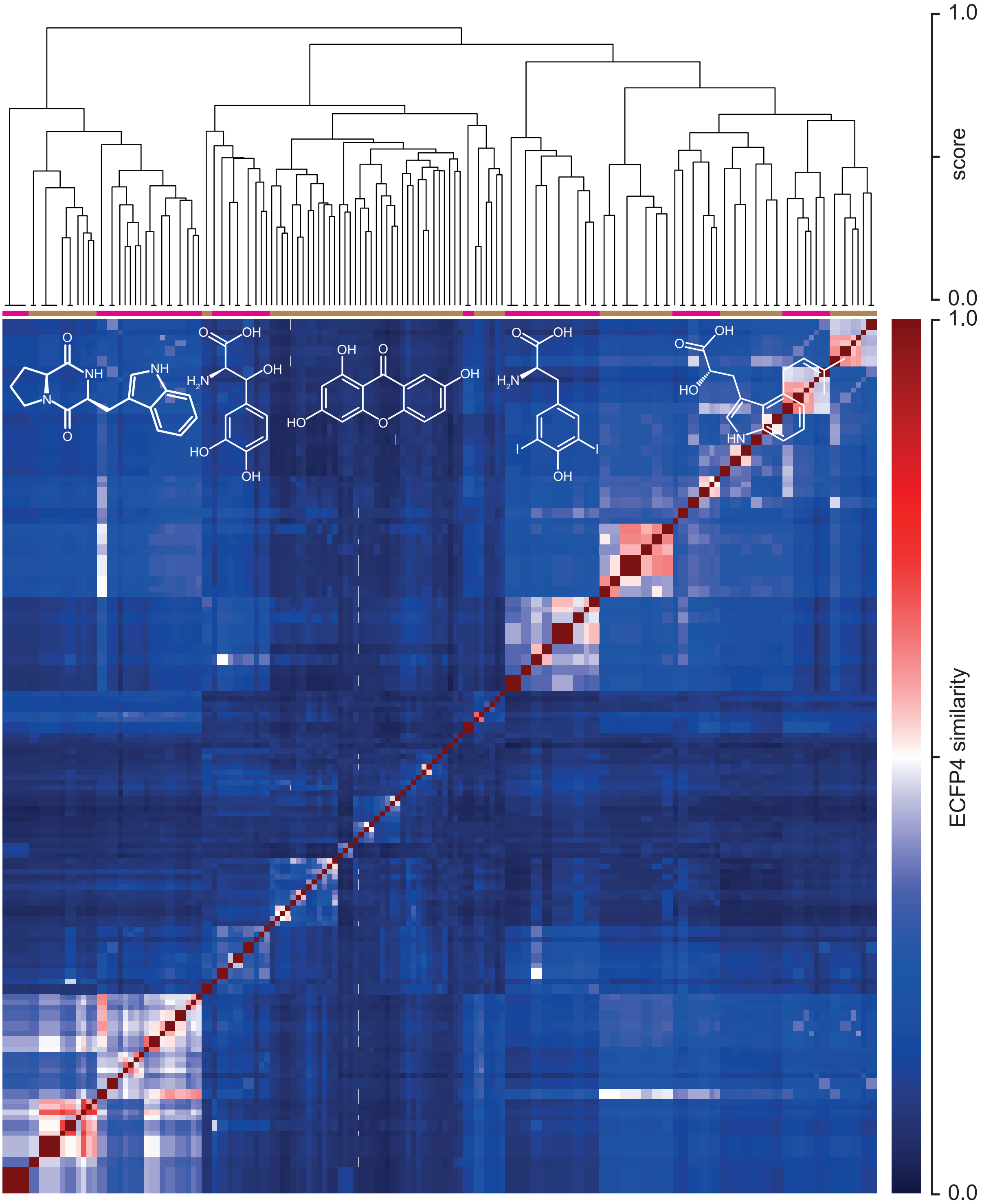


Figure 4: Hierarchical clustering of the substrates extracted from PrenDB. Individual clusters derived from the dendrograms are annotated with an image of the corresponding cluster representative. Magenta and brown bars indicate 14 detected clusters. Black horizontal lines on the leaves of the dendrogram indicate the number of molecules grouped together.

with the DMATS bias in the literature, eventually leads to strongly indole-biased data.

Figure S1 shows the knowledge about prenyltransferase reactions, as digitalized and stored within PrenDB, in terms of catalyzed transformations – combinations of a particular substrate and an enzyme – and the corresponding yield achieved. In the top left corner of the matrix, transformations of the most abundant substrates (tryptophan, tyrosine, cyclic dipeptides and their derivatives, respectively) together with the most promiscuous enzymes (7-DMATS,

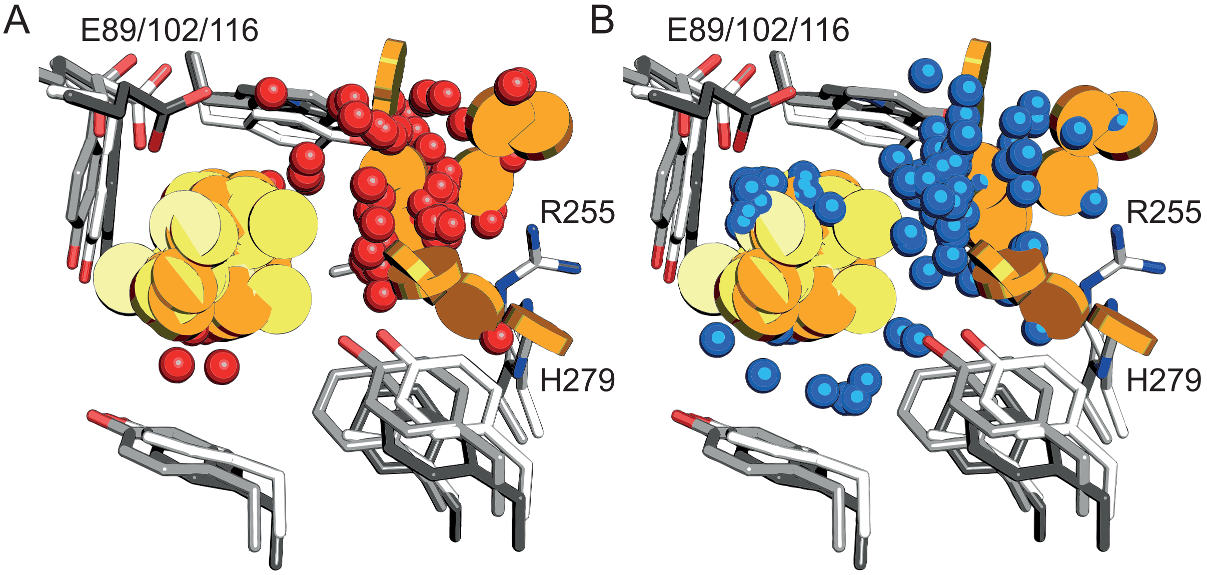


Figure 5: Molecular features extracted from poses of the selected virtual hits. **A** Red spheres indicate H-bond acceptors and the yellow discs aromatic moieties. The majority of acceptor functionalities can be found around the basic residues R255 and H279. **B** Blue spheres indicate H-bond donors. They are located around the highly conserved glutamate (E89/102/116) and in the vicinity of backbone carbonyls (omitted for clarity).

FtmPT1, CdpNPT, SirD and FgaPT2) can be found. At the same time, the matrix is sparse, *i.e.*, contains a lot of blanks. This sparsity is the result of the availability of data and thus represents the research focus in the prenyltransferase field in the past. It presents a challenging starting situation for model building.

2.1.2. Repitopes

For each of the 665 cataloged reactions – each defined as a unique triplet of a substrate, product and enzyme – a repitope (*reactive epitope*; *cf.* Materials and Methods for complete definition) was extracted using the reactive atom detection and repitope reconstitution routines of the algorithm developed in this work (*cf.* Materials and Methods). Each repitope comprises four reconstitution depths (from 2 to 5 bond distances). Over all repitope depths, 276 unique repitopes, defined by their unique SMARTS string, were extracted. A SMARTS string is a one-dimensional encoding of chemical substructures and an efficient way to store a complete definition of each repitope (*cf.* Materials and Methods). A reconstitution depth of 5 delivered the largest contribution to the repitope space with 135 (49 %) members. This is consistent with expectation, as larger depths will lead to more diverse descriptions. Depths 2, 3 and 4 account for 25 (9 %), 69 (25 %) and 94 (34 %) repitopes, respectively. Interestingly, the sum over all amounts of each reconstitution level is greater than the total number of unique repitopes. This means that distinct combinations of the substrate molecule, its reactive atom and the depth level are not mutually exclusive, thus resulting in duplicate entries. **Figure S2** shows the distribution of repitopes for the various depth levels of reconstitution, underlining the relationship between repitope size and diversity. Although the largest class of repitopes contributes the most to repitope space, also smaller, more general and ambiguous repitopes are among the top ten most frequent repitopes: Substructures of tyrosine, benzene and indole moieties are at ranks 3, 5, 6, 7 and 9, respectively, whereas complete and extended indole and tyrosine moieties are at ranks 1, 2, 4 and at the bottom, 8, 10 and 11.

2.2. Prediction of novel substrates via a multi-step screening procedure

In a sequential application of virtual screening tools (**Figure 2**), beginning with prenylation prediction through repitopes stored in PrenDB and concluding with docking into three prenyltransferases with known crystal structures (FgaPT2, FtmPT1 and CdpNPT), 38 virtual hits were selected through the following procedure: i) A compound was considered as a virtual hit if any PrenDB repitope could be found within its molecular framework at least once. **Table S2** shows the number of repitopes matching a particular hit, with a high repitope hit rate indicating promiscuous compounds, *i.e.*, molecules that are classified as substrates of multiple enzymes. Using repitopes based on a reconstitution depth of three, 168,906 compounds were selected in this first step. ii) Comparison of molecular properties with those of known substrates and removal of molecules outside the respective ranges (**Table 2**), reduced the number of virtual hits to 90,559. By going beyond 1D and 2D molecular descriptors and ensuring that the iii) three-dimensional shape (judged by a high score in the OEChem shape congruency tool, *cf.* Materials and Methods) matched between putative and known substrates, lead to a selection of 451 compounds. This repitope-, property- and shape-based determination of prenylation potential of the selected compounds was further condensed by the iv) docking results: For each compound, an optimal enzyme structure for docking was selected based on a compound's structural overlap with the co-crystallized substrate. The amount of this overlap was quantified by the same shape congruency methodology mentioned above, but was automatically invoked from within the docking application HYBRID (*cf.* Materials and Methods). The generated poses, from which 38 molecules were selected for experimental validation, show a distinct geometrical consensus of the key interactions with the enzymes (**Figure 5**): first, polar interactions with the general base E89/102/116, second, occupation of the apolar indole-subpocket and H-bond interactions in the vicinity of the opening of the active site – residues H279 and R244.

2.3. Novel substrates for prenyltransferases FgaPT2, FtmPT1 and CdpNPT

In order to assess the predictive performance of our virtual screening, the 38 potential substrates were tested as prenyl acceptors in enzyme assays with the tryptophan prenyltransferase FgaPT2 and the two tryptophan-containing cyclic dipeptide prenyltransferases FtmPT1 and CdpNPT. The selected substances clearly differ structurally from the substrates for the DMATS prenyltransferases reported previously. (7,10)

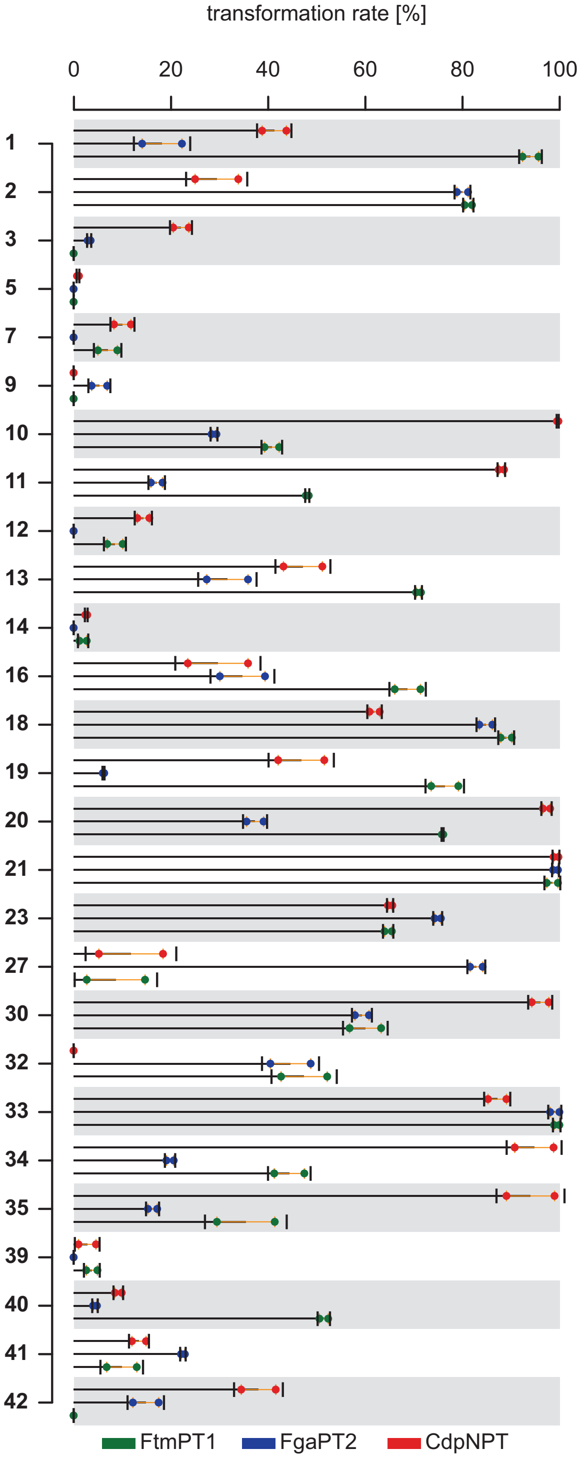


Figure 6: Transformation rates of virtual hits relative to L-tryptophan obtained for the three examined enzymes FtmPT1, CdpNPT and FgaPT2. Horizontal bars indicate the mean, vertical bars the standard deviation, orange interval the standard error of the mean and colored circles the data points, respectively.

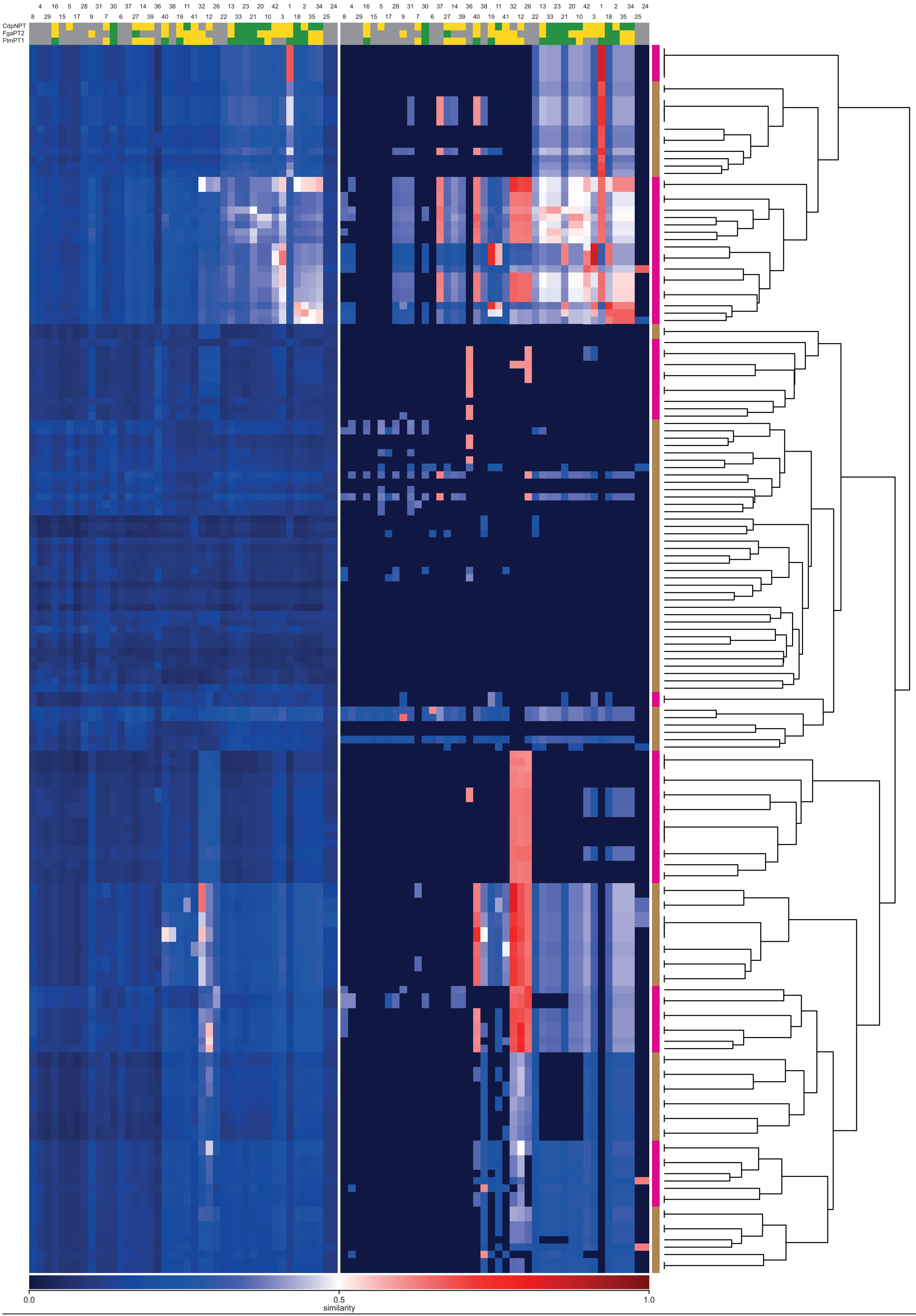


Figure 7: Similarity matrix between the selected compounds and known substrates for prenyltransferases extracted from PrenDB. **Left** ECFP4 fingerprint similarity. **Middle** RedFrag scores calculated with ECFP4 fingerprints. Color coding (top), green, yield > 50%; yellow, yield between 1 and 50 %; gray, no transformation. **Right** magenta and brown bars indicate 14 detected clusters. Black vertical lines on the leaves of the dendrogram indicate the number of molecules grouped together.

The reaction mixtures were analyzed with LC-MS in order to detect the formation of prenylated products. As shown in **Table S2** and **Figure 6**, 23 of these substances were accepted by FtmPT1, 22 by FgaPT2 and 25 by CdpNPT. In relation to the number of hits selected from our virtual screen, this corresponds to a hit rate of 60.5 % for FtmPT1, 57.9 % for FgaPT2 and 65.8 % for CdpNPT. Product yields of more than 50 % were observed for 12 substrates with FtmPT1, 7 with FgaPT2 and 10 with CdpNPT, respectively. The prenylated products can be detected in a straightforward manner as signals in their corresponding mass spectra: Their $[M+H]^+$ ions are shifted by 68 Daltons relative to their educts. Overall, we thus obtained high hit rates and yields higher than 50 % in case of 29 reactions (25 % of all attempted reactions).

2.3.1. Similarity analysis of known substrates and selected compounds

In order to assess the novelty of the 38 selected compounds, the similarity with the substrate space cataloged within PrenDB (167 substrates) was calculated and visualized by generating a similarity matrix based on the ECFP-fingerprint-based Tanimoto similarity (**Figure 7** left panel). The matrix shows an overall low similarity score between our selection and the known substrate space. This points towards the potential to access truly novel substrate space by employing repitopes. Of note, the similarity is higher in columns corresponding to compounds that were successfully prenylated in our assays by at least two of our test enzymes. The right panel of **Figure 7** shows similarity scores as calculated by our in-house fragment-based method RedFrag (12): In contrast to ECFP4, RedFrag compares the fragmental composition of molecules and the 2D arrangement of fragments. RedFrag accentuates the commonalities and differences between known substrates and our selections. Compound **1**, a tryptophan-homo-proline-diketopiperazine (94.1 % yield on FtmPT1), shows high similarity scores with tryptophan, its indole-core derivatives and, expectedly, with tryptophan-tryptophan-, tryptophan-alanine-, tryptophan-glycine- and tryptophan-proline-diketopiperazines from clusters 1, 2 and 3, respectively. Of note, RedFrag emphasizes the similarity of compound **1** and cluster 3 based on the presence of the indole scaffold. In contrast, ECFP4 emphasizes the dissimilarity of this compound originating from the absence of the diketopiperazine motif in the same cluster. Compounds **12**, **26** and **32** show high similarity with substrates from cluster 3 (also 9, 10 and 11). Compounds **12** and **32** are regioisomers of a brominated tryptophan derivative. They show distinctly different yields: 14.4 and 8.5 % for **12** on CdpNPT and FtmPT1; 47.5 and 44.7 % for **32** and FtmPT1 and FgaPT2, respectively. It is evident that the position of the bromine atom has a major impact on the role of such compounds as substrates. The influence of regiochemistry of indole-core substitutions or single-atom-replacements at this core is further exemplified by compound **26**. Its benzothiophene moiety (replacing the nitrogen atom in an indole by a sulfur atom) is not accepted as a substrate by any of the three test enzymes.

Selected compounds with low similarity, but remarkable yields, indicate novel substrate classes or motifs: Compound **16** shows a good yield in FtmPT1 and moderate yields in FgaPT2 and CdpNPT (68.8, 34.8 and 29.7 %, respectively). Its conjugated indole-4-imidazolin-2-one motif has no similar counterparts within the known substrate space. This is also true for compound **30** (yields: 60.1, 59.4 and 96.1 %, respectively) and its benzylated hydroxyl-indole structure. Compound **27** – a pyrimidine-indole – shows excellent yield in FgaPT2 (82.9 %). Further examples with high yields but RedFrag similarity scores lower than 0.6 are compounds **11**, **13**, **20**, **23** and **33**.

2.3.2. Structure elucidation of the products of compounds 21 and 30

To investigate at which position within a given substrate the prenylation occurred, we carried out exemplary FtmPT1 incubations with two indole derivatives, indole 3-ethylamine ethylsufonamide (**21**) and 6-benzyloxyindole (**30**), which were very well (98.6 % yield) and moderately (60.1 % yield) accepted by this enzyme, respectively (**Figure 8** and **Table S2**). As shown in **Figure S3A**, a single dominant peak

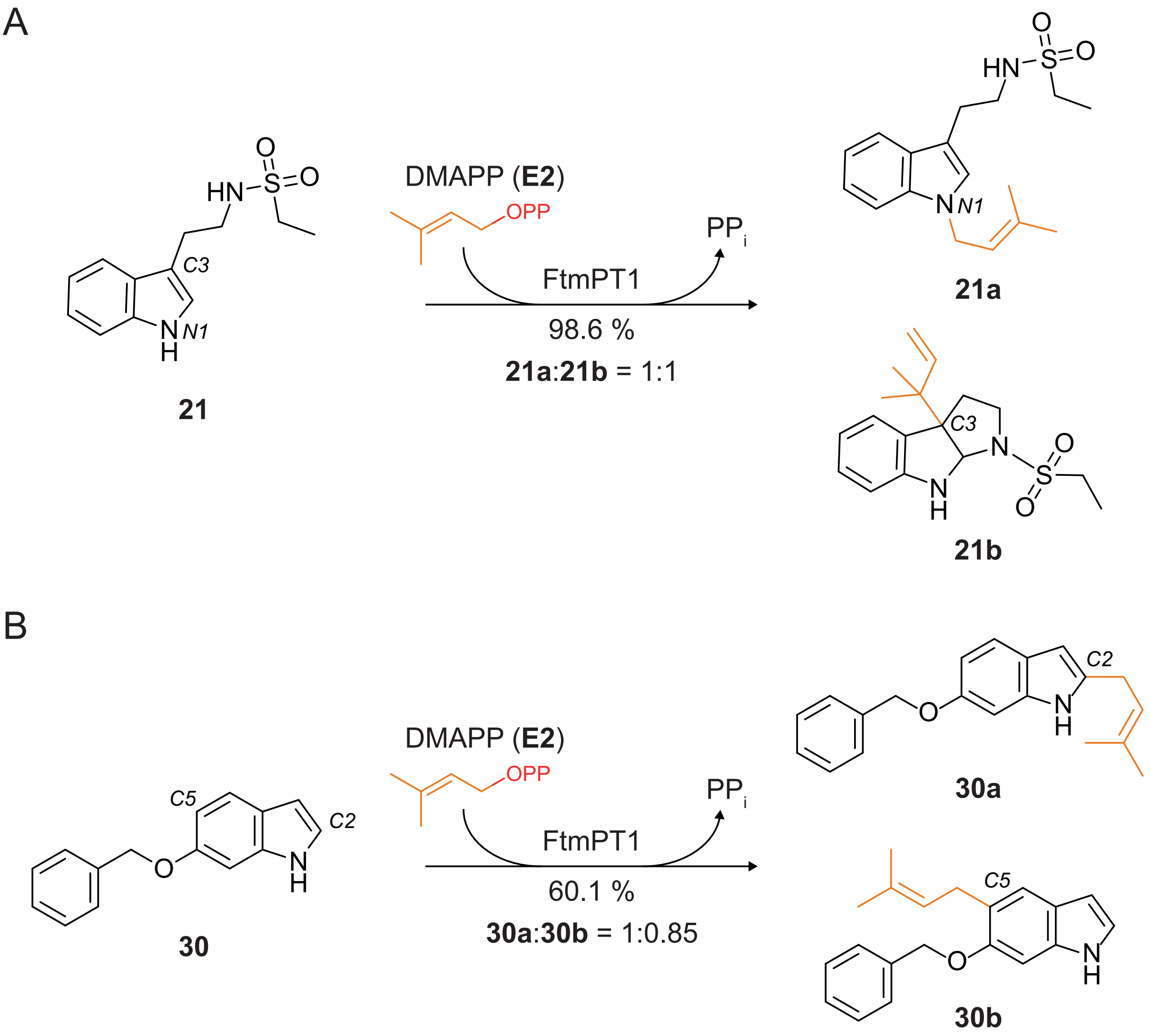


Figure 8: Transformations of virtual hits **21** and **30** by FtmPT1. **A** Indole 3-ethylaminine ethylsulfonamide (**21**) was regularly prenylated at position *N1* leading to **21a** and reversely prenylated at position *C3* with a simultaneous formation of a 6/5/5 fused ring system (**21b**). Typically for *C3*-prenylation at indole substructures, a dearomatization and intra-molecular cyclization accompanies the prenylation reaction. **B** 6-benzyloxyindole (**30**) was regularly prenylated at positions *C2* and *C5*.

was observed in the LC-MS chromatograms of the incubation mixtures, which were isolated on a Multospher 120 RP-18 column for structure elucidation. ^{1}H NMR data revealed, surprisingly, the presence of two compounds in each reaction mixture. **21a** and **21b** originated from **21** in a ratio of 1:1 from the reaction mixture, and **30a** and **30b** in a ratio of 1:0.85 from that of **30**. Further purification resulted in four pure products. Through NMR and MS analyses, the structure of **21a** was subsequently elucidated as a regularly *N1*-prenylated derivate. The second product, **21b**, was identified as a reversely *C3*-prenylated derivative with a simultaneous cyclization of *C2* of the indole with the nitrogen atom of the side chain located at *C3*, resulting in the formation of a 6/5/5 fused ring system. Compounds **30a** and **30b** were proven to be regularly *C2*- and *C5*-prenylated derivatives, respectively. These results unequivocally proved specific prenylations at the indole ring without or with additional modifications such as cyclization. Detailed studies of the relationships of enzymes, substrates listed in **Table S2**, and their products are under further investigation.

A comparison of the elucidated structures of the products **21a**, **21b**, **30a** and **30b** with the PrenDB-predicted prenylation sites of their corresponding educts reveals that the prenylation site was correctly predicted in two of four cases. However, the responsible enzyme, FtmPT1, was only proposed for the prenylation of **21** to yield

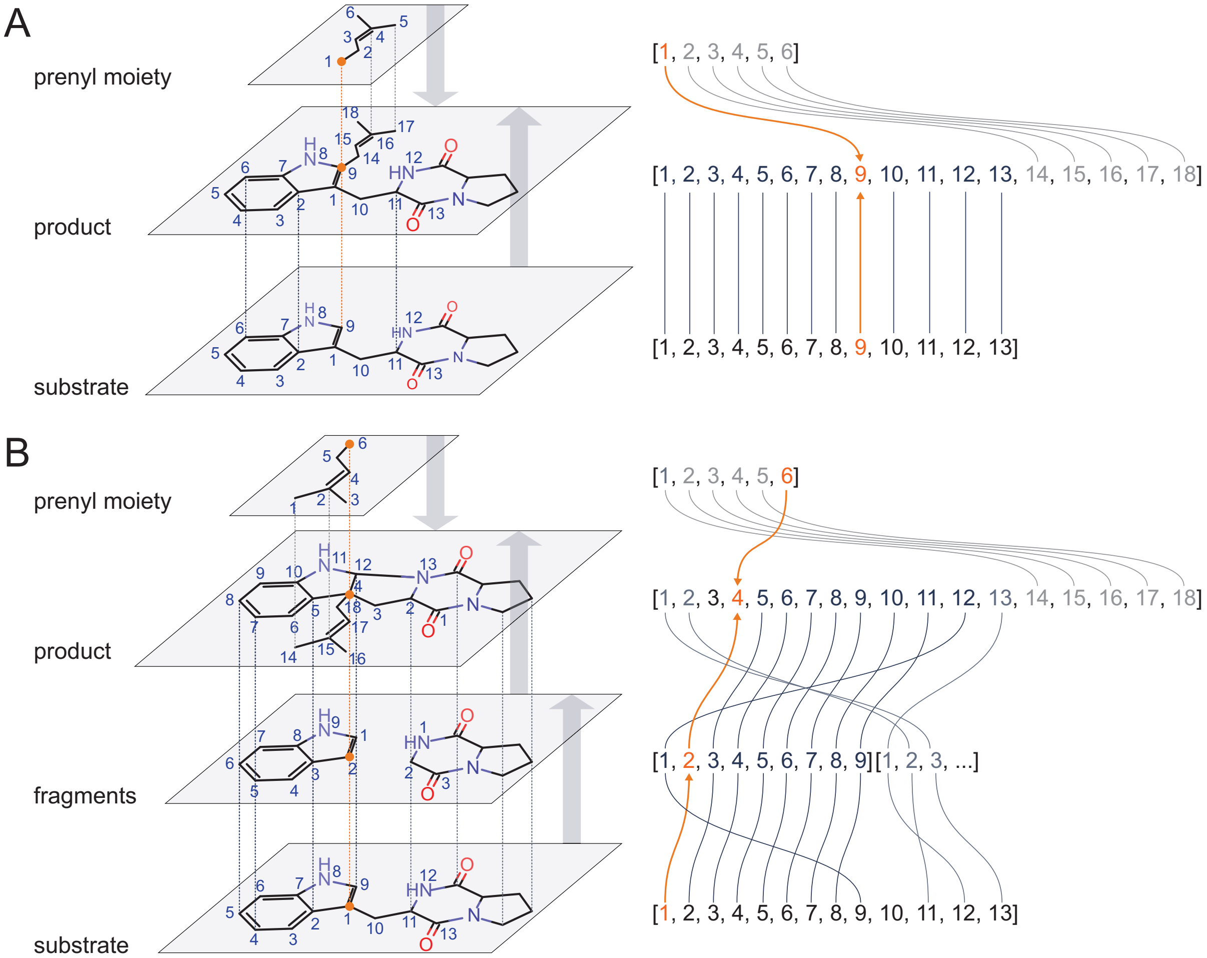


Figure 9: A Substrate-based subgraph isomorphism. The substrate structure matches the product as a whole. The intersection of atom overlaps between substrate and prenyl moiety delivers the reactive atom (orange arrows, index 9). **B** Fragment-based subgraph isomorphism. Substrate structure is fragmented into smaller epitopes preserving the substrate-fragment atom matchings. By matching fragments onto the product and analyzing the intersection with the prenyl moiety, the reactive atom can be found within the structure of the substrate (orange arrows, index 1).

21a. In case of **30**, the product **30b** was predicted to originate from the enzymes CdpNPT or FgaPT2.

3. Discussion and conclusions

This study demonstrates the power of systematically organizing and analyzing diverse and disparate experimental enzymatic data by means of chemoinformatic methods. Besides a comprehensive repository of the existing knowledge about prenyltransferase reactions, the determination of repitopes allowed us to predict novel substrates that are distinctly different from the ones that have been identified previously, both natural and synthetic. Moreover, we achieved an overall high hit rate of 71 % in terms of molecules that were accepted by at least one prenyltransferase. However, it has to be noted that the repitopes stored in PrenDB are not yet accurate enough in all cases to precisely predict the correct enzyme and/or the correct reactive atom. This shortcoming is presumably correlated with the comparatively small number of instances in the database. Although the existing body of literature clearly represents a considerable experimental effort, chemistry and the biochemical reactivity of enzymes are so diverse that even higher numbers of substrate-enzyme-product triplets would be necessary to obtain more complete repitopes that also account for the different reactivity of certain substructures. The chemoinformatic strategy that we employed in this work is certainly flexible enough to accurately model more fine-grained patterns.

At the same time, a database such as PrenDB can provide excellent help in determining which reactions and substrates would be worthwhile to

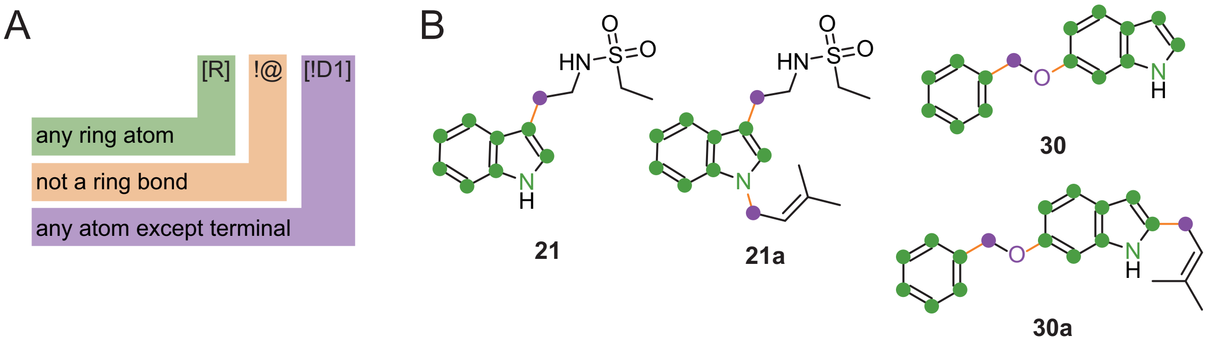


Figure 10: **A** Fragmentation rule expressed as SMARTS string. The rule consists of two atom definitions and a bond definition. Enclosed in square brackets (green and purple) are atoms connected by any bond type except a ring bond (orange). The atom on the left-hand side can be of any type but must be member of a ring system (hydrogen atoms are excluded indirectly as they are not allowed to form ring systems). On the right-hand side, the atom must not be a terminal atom (hydrogen atoms are excluded indirectly as they are always terminal). *D1* represents atoms with only one neighbor. **B** Breakable bonds (orange) as defined by the fragmentation rule and atoms color coded based on the scheme in **A** for substrates **21** and **30** and their corresponding prenylation products **21a** and **30a**.

test next. On a basic level, one could simply be guided by the number of reactions already described for each enzyme and focus on the underrepresented ones. But also, more sophisticated approaches can be envisioned: Enzyme phylogenetic trees could be based not on amino acid sequence, but on substrate similarity. Further exploration would thus focus on filling in the *missing links*. Ultimately, such strategies might merge with machine learning approaches, where the algorithm itself would suggest which enzyme-substrate pairs to test next based on the maximum information gain of each investigation.

Yet, despite some shortcomings, our database and resulting prediction algorithm are already useful for correctly predicting a large number of substrates and thereby aiding in the creation of novel chemical matter. The imprecisions could also be taken as a strength in this context, as they allow for serendipitous discoveries, *e.g.*, the reverse prenylation of **21** to **21b** by FtmPT1.

Lastly, it has to be emphasized that the concept of repitopes and their fragment-based determination can easily be extended to other enzymatic reactions. The automatic processing of potentially large numbers of reactions and the concomitant conversion into the reaction principles (*i.e.*, repitopes) will lead to facile systematizations and gain of knowledge from the analyses of the emerging data.

The high hit rates (58 - 66 %) for each enzyme and the fact that one fourth of the reactions had a yield of 50 or more percent demonstrates the excellent performance of our knowledge-based repitope approach. The combination of PrenDB and its ligand-based approach with protein-structure-based tools such as docking therefore seems to constitute a powerful combination of strategies. Furthermore, these results prove the potential usefulness of the tested enzymes to produce prenylated derivatives.

4. Materials and Methods

The prenylation reaction as conducted by the enzymes of the DMATS superfamily formally corresponds to a substitution reaction occurring on carbon, oxygen and nitrogen atoms of small metabolites through the transfer of small apolar moieties (denoted as dma [short for dimethyl allyl] in the following example). The leaving group is always a pyrophosphate (PP_i) and a formal proton accepted by a general base. The reaction can be written in a symbolic way (SMIRKS notation (13), see below): The atoms taking part in the chemical transformation are arranged in a one-line notation showing the bond cleavages and formations.

[C,O,N:1][H:2] + dma[C:3][O:4]PP_i + [gB:5]

→

[C,O,N:1][C:3]dma + [O:4]PP_i + [gB:5][H:2]

Square brackets enclose individual atoms and adjacent atoms are taken to be linked by covalent bonds. Letters denote elements, gB is the general base and PP_i pyrophosphate, commas represent a logical *OR* and numbers are arbitrary labels to allow for unambiguous tracking of each atom. In the above example, it can be seen that the hydrogen atom with label 2 ([H:2]) is substituted by the dma group and moves from its adjacent carbon, oxygen or nitrogen atom (labeled 1) to the general base (label 5). **Figure 1A** illustrates this general transformation in a 2-dimensional way exemplarily for the reaction between brevianamide F (**E1**) and DMAPP (**E2**) catalyzed by FtmPT1. With this symbolic notation and common chemoinformatic tools in hand, it is possible to virtually transform, for example, any carbon atom [C:1] bearing a hydrogen atom [H:2] (**GA** in **Figure 1C**) into the prenylated product **GP,** at the same time generating a protonated general base **PGB** and pyrophosphate (**PP**) as byproducts. Although feasible *in silico*, a chemical transformation based solely on the position of the reactive atom is unreasonable and ambiguous – in reality, only carbon, nitrogen and oxygen atoms located within the correct atomic surroundings can undergo prenylation. Thus, the entire molecule, or at least a crucial motif within it, is necessary to completely characterize a reactive environment., We call such a set of atoms consisting of the reactive atom – to which the transferred moiety will be attached – and its neighboring atoms a *reactive epitope* (repitope for short). In the case of the transformation of **E1**, the corresponding repitope is shown in **Figure 1D**. The specification of the carbon atom can now be extended to its full repitope notation:

[cH0][nH]**[cH1:1]**[cH0]([CH2])[cH0]

Where small letters denote membership of an atom in an aromatic system, HX states the presence of X adjacent hydrogen atoms and parentheses indicate branching of the molecular framework. From this notation, it can be concluded that the reactive atom is aromatic and is bound to one hydrogen atom; its direct neighbors are an aromatic nitrogen and another aromatic carbon without any attached hydrogen atoms. The second neighbor shell consists of two aromatic and one aliphatic carbon atoms. This convenient one-line notation of chemical environments is called SMARTS (one-line molecular patterns) (13) and is widely used in the field of chemoinformatics – especially for substructure searches: Atoms, their properties and binding characteristics are encoded with alphanumeric characters. Multiple molecule SMARTS together with the information about bond breakage and formation yield the SMIRKS of a reaction. With a repitope, such as the one described above, and the enzymatic transformation encoded in SMIRKS notation, it is possible to virtually transform any substrate molecule into its corresponding prenylation product or to easily search for putative substrates by invoking substructure-based virtual screens in publicly available vendor databases.

The corresponding SMARTS notation for each repitope could in principle be deduced by hand given the chemical structures of the substrate and product molecules. In order to achieve an efficient handling of several hundreds of enzymatic transformations – with only the 2D structures of substrate, product, and the transferred moiety as input – an automated procedure for the extraction of transformation SMARTS, and thus repitopes, appears to be as indispensable as it is difficult to accomplish. A fully specified repitope requires the knowledge of the reactive atom as well as its surroundings. Repitope deduction can be accomplished by applying subgraph isomorphism-based algorithms followed by the reconstitution of the chemical environment. Both steps – reactive atom perception and repitope reconstitution – will be described in detail below. All coding was done in python. For chemoinformatic calculations, the python wrappers of the RDKit (15) library were utilized. Fingerprint-based similarity calculation was carried out with the OEChem toolkit. (16)

4.1. Perception of the reactive atom

In case of a simple linear substitution reaction as depicted in **Figure 1A**, the reactive atom can be found by mapping the molecular structure of the substrate molecule onto the molecular structure of the product. For non-symmetric molecules, this leads to a unique match with an atom-to-atom correspondence between substrate and product. As the number of atoms in the product is always greater than the number in the substrate, the substrate is a substructure of the product, *i.e.*, its

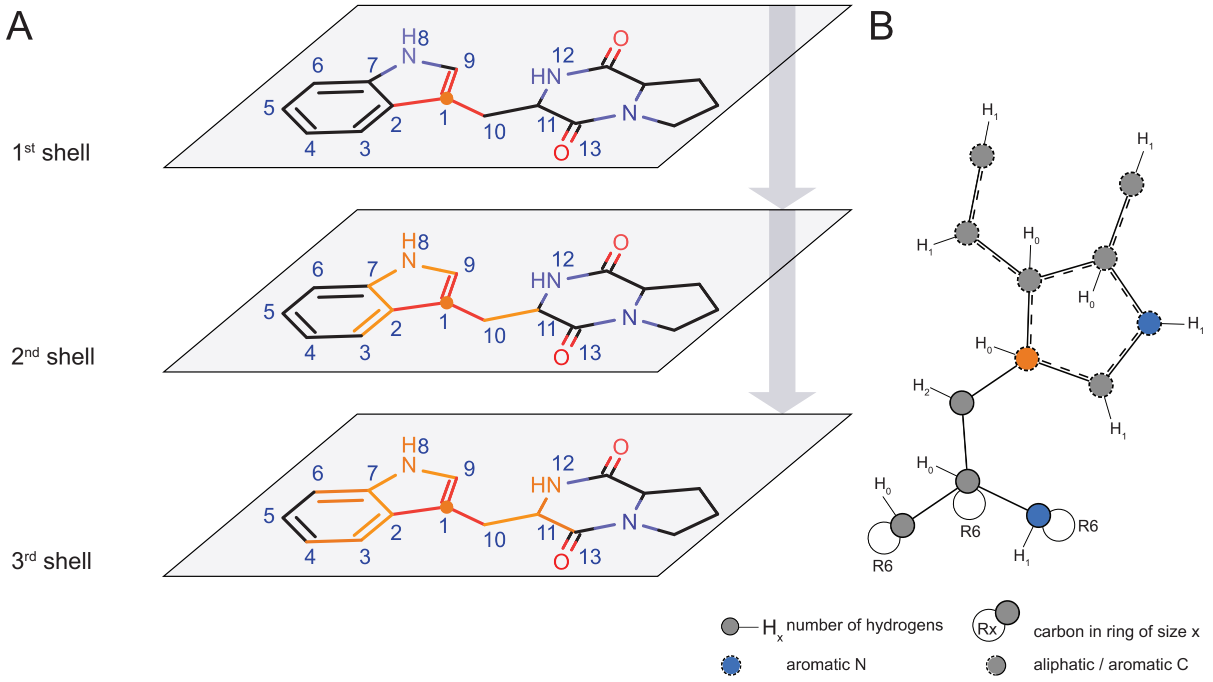


Figure 11: **A** A repitope is generated by sequentially rebuilding the substrate molecule shell by shell with the reactive atom (index 1) as anchor point. Each iteration adds another neighbor shell to the repitope resulting in a fully defined depth-3 repitope as depicted by SMARTSviewer (**B**).

complete molecular skeleton can be found within the one of the product. With the same approach, the atom-to-atom correspondence between the transferred moiety (*i.e.*, the prenyl group) and the product molecule can be obtained. The intersection of the atom-to-atom matched sets of the substrate and the transferred moiety consists of only one atom – the reactive atom (**Figure 9A**). If, however, the enzymatic transfer of a moiety is accompanied by a subsequent (or concerted) rearrangement of the molecular skeleton of the product (*e.g.*, a cyclization), the substrate cannot be considered to be a direct substructure of the product anymore. Thus, the reactive atom can no longer be determined through the atom wise substrate-to-product mapping as described above. In such a case, a possible strategy for establishing a substructure correspondence, *i.e.*, a subgraph isomorphism, would be to weaken atom or bond type matching criteria. The resulting atom-to-atom correspondences are allowed to be more general in that way but are often ambiguous at best. In order to circumvent this problem, we assumed that, although the entire substrate may undergo dramatic changes in its molecular skeleton, smaller structural motifs (molecular fragments) remain unaffected by such transformations and can therefore still be unambiguously mapped onto the substrate structure *before* and *after* the reaction. **Figure 9B** illustrates the consecutive steps in this *fragment-based* substructure isomorphism approach. In contrast to the aforementioned subgraph isomorphism based on the entire substrate structure, an additional fragmentation step has to be performed: Breakable bonds, (bonds connecting ring systems with other ring systems or with acyclic motifs [*cf.* **Figure 10A** and **B** for exemplary fragmentations and the breakable bond definition)] are cleaved, leading to a set of molecular fragments each substrate is constituted of. An intermediate filter step ensures that very small fragments (single atoms, linker moieties, terminal groups) are not considered further. The remaining fragments are mapped onto the product structure and their atom-to-atom correspondence is investigated for an intersection with the atom mappings of the transferred moiety and the product molecule. By preserving the atom-to-atom correspondences between the substrate molecule and its fragments, the reactive atom can be identified by finding the intersecting atom of one of the matching fragments in the substrate.

4.2. Reconstitution of the reactive epitope

As already mentioned, knowledge of the reactive atom alone is only of limited use for substructure searches or virtual transformations, because both methods yield ambiguous results when only a single atom is given as input. It is therefore necessary to rebuild the chemical environment of the reactive atom in order to obtain a description of a particular transformation that has discriminative power. To obtain such a description, the reactive atom is augmented with additional atoms from its first, second, third, (etc.) neighbor shells (**Figure 11**), *i.e.*, by traversing the atomic neighborhood of the reactive atom up to a fixed distance (*i.e.*, number of bonds). The traversed atoms are then extracted as a molecular subset and converted into a regular molecular object and, eventually, a SMARTS string. Different depths of reconstitution lead to either small and unspecific repitopes ($d = 1$) or larger and more stringent ones for large depths ($d > 3$). *In extremis*, at the largest possible depth, the repitope becomes identical to the molecule itself. Thus, a balance must be found and the most useful repitopes are able to represent reasonable chemical environments for a particular reaction, but still allow for certain flexibility and diversity in retrieving putative substrates.

4.3. Database of prenylation reactions

With the algorithmic tools to deconstruct a given transformation catalyzed by a prenyltransferase into a reaction SMARTS and the corresponding repitopes in hand, investigation of as many transformations as possible can readily be conducted. Hence, we decided to create a database (PrenDB) storing the known transformations in an efficiently browsable and queryable manner. For this purpose, a literature search was performed to extract substrates, products, enzymes and available meta data (such as kinetics and yields) from 44 publications – full

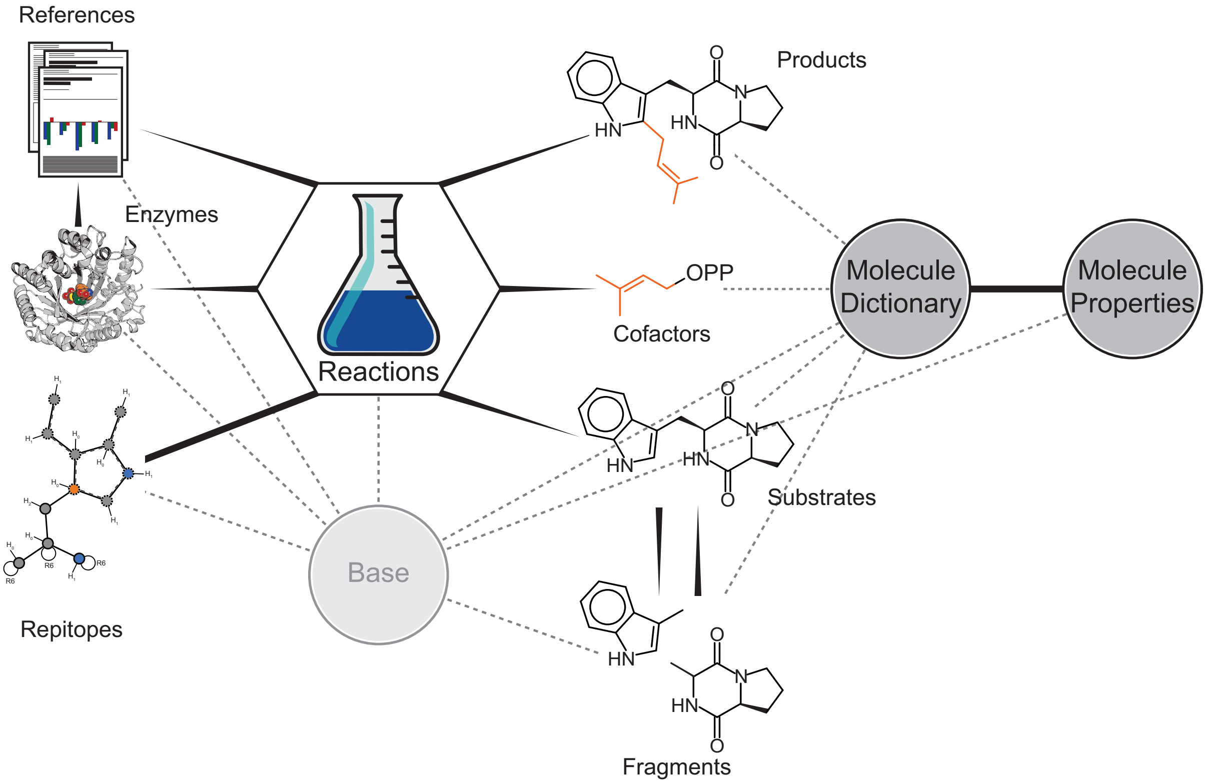


Figure 12: Design of PrenDB. The database tables are related to each other in a one-to-one (reactions and repitopes), one-to-many (substrates and reactions) or many-to-many (substrates and fragments) relationship, reflecting their real-world correspondence. The central reaction dictionary holds the necessary data to encode a reaction based on substrate, product and cofactor molecules, the enzyme and the resulting repitope. A reference table is added in order to supplement the database with meta-data and enhance its usability. Dashed lines indicate abstract inheritance. Solid wedged lines represent one-to-many relationships, *e.g.*, a molecule can act as substrate in as many reactions as an enzyme. A repitope belonging solely to one particular reaction is indicated by a straight solid line (one-to-one relationship).

articles, reviews and communications – across 17 journals. Each enzymatic reaction is represented by the SMILES strings of the product and substrate molecules, combined with the preferred name of the involved enzyme. The advantage of using SMARTS is that each reaction can be visualized and processed with common chemoinformatic software. Furthermore, each reaction entry contains multiple repitopes, generated with the aforementioned algorithm and different environmental depths (2 to 5 bonds around the detected reactive atom). This reaction table (a *table* is a collection of database entries that are semantically equal) is supported by and connected with further tables holding meta data extracted from the literature and/or calculated with chemoinformatic tools (**Figure 12**). The molecule dictionary table comprises all small molecules involved in the reaction: substrates, products, transferred moieties (such as DMAPP or benzyl pyrophosphate), and fragments. Additionally, each entry comes with a molecular properties table, where basic physico-chemical properties can be looked up. The reference table contains the literature used for data extraction together with hyperlinks to articles and entries on PubMed and UniProtKB. PrenDB can be browsed and extended with python scripts bundled with the algorithms for repitope generation described in this work (or more conveniently via a web interface) in a straightforward manner. Because of access speed and portability considerations, we decided to use the sqlite3 backend as the underlying database architecture and the Django python package for middleware and frontend.

4.4. Virtual screen for putative substrates of prenyltransferases

In order to predict novel substrates for transformation by a prenyltransferase, a multi-step screening process was carried out with a subset of the ZINC database (17) which stores commercially available small molecules in ready-to-be-processed formats (**Figure 2B**). First, the

Table 2: Range limits of physico-chemical properties derived from the substrate space stored in PrenDB.

physico-chemical property	min	max
molecular weight [Da]	140	515
number of heavy atoms	11	37
number of carbon atoms	8	27
number of heteroatoms	1	10
number of chiral centers	0	5
H-bond acceptors	0	6
H-bond donors	1	6
number of atoms in a ring system	6	25
number of rotatable bonds	0	6
number of rigid bonds	9	40
XLogP	-3.94	3.76
minimal solubility attribute	poorly[a]	
2D polar surface area [Å^2]	20.0	1125
removal of known aggregators	true[b]	

[a] Solubility categories (insoluble, poorly, moderately, soluble, very, highly) are derived from reparametrized atom-types from the XLogP algorithm. (16)

ZINC clean leads database, with a total of 5.1 million compounds, was filtered for the presence of any of the extracted repitopes from PrenDB. The repitope depth was 3. The screening was carried out as substructure searches using the python wrappers of the OEChem toolkit. (16) Second, compounds were filtered utilizing the MolProp toolkit. (16) Only compounds with physico-chemical properties within the range spanned by known substrates cataloged in PrenDB were allowed (**Table 2**). The remaining compounds were subsequently submitted to the shape congruency analysis based on the OEChem API. (16) In short, for each compound a low-energy conformer was generated and its 3-dimensional overlay with each known substrate was optimized. Compounds with an overlay score greater than 0.9, reflecting excellent 3D-shape matching, were allowed for the next step. Fourth, the remaining compounds were docked into the three most promiscuous prenyltransferases for which a crystal structure has been determined, (FgaPT2, FtmPT1 and CdpNPT; PDB codes 3I4X, 3O2K and 4E0U, respectively) employing the multi-target HYBRID (18) engine: For each compound, up to 200 conformers were generated with OMEGA (Hawkins *et al. J. Chem. Inf. Model.*, **2010**, *50*, 572-584.). The ensemble of conformations of each molecule was then overlaid with the co-crystallized ligand in each of the three selected crystal structures in order to determine the best suited enzyme for the following exhaustive docking. The method for overlaying conformers is built directly into the HYBRID engine and is based on the same methodology as implemented in the OEChem API and the ROCS application (Hawkins *et al.*, *J. Med. Chem.*, **2007**, *50*, 74-82.). For the actual docking step – translational and rotational optimization of a compound conformer within the binding site of the protein – HYBRID scores for a given protein-ligand complex were calculated based on the shape and electrostatic complementarity of the ligand and protein's binding site (**Figure S4**). Shape and electrostatic features are represented by Gaussian potentials. During optimization, the overlap between ligand and protein features is maximized. After docking, Calculated poses were visually inspected in order to remove those that form improbable interactions that are not sufficiently penalized by present-day scoring functions and the selected compounds were acquired from their respective vendors and experimentally tested.

5. Experimental validation

5.1. Chemicals, bacterial strains and culture conditions

DMAPP was synthesized according to the method described for geranyl diphosphate reported previously. (19) The 38 tested substrates were purchased from Enamine Ltd, Kiev, Ukraine; ChemBridge Corporation, San Diego, USA; MolPort, Riga, Latvia; Vitas-M Ltd, Apeldoorn, Netherlands; Mcule, Inc, Budapest, Hungary.

Escherichia coli strains XL1 Blue MRF' (Stratagene, Heidelberg, Germany) and *E. coli* BL21 (DE3) (Invitrogen, Karlsruhe, Germany) were used for protein overproduction. The strains with expression plasmids were cultivated in lysogeny-broth (LB) or Terrific-Broth (TB) medium at 37 °C with 50 $\mu g \cdot ml^{-1}$ carbenicillin or 25 $\mu g \cdot ml^{-1}$ kanamycin as selection marker. Overproduction of FtmPT1 with pAG012, FgaPT2 with pIU18 and CdpNPT with pHL5 were carried out as reported previously. (20,21,22)

5.2. Enzyme assays with recombinant proteins

In the assays to determine the acceptance of the different substrates, the enzyme reaction mixtures contained 50 mM Tris-HCl, pH 7.5, 10 mM $CaCl_2$, 2 mM DMAPP, 2-7.5 % (v/v) glycerol, 1-2 % (v/v) dimethyl sulfoxide (DMSO), 1 mM aromatic substrate and 0.4 $mg \cdot ml^{-1}$ purified recombinant protein in a volume of 100 µl. The reaction mixtures were incubated at 37 °C for 16 h and terminated by addition of an equal volume of methanol. The reaction mixtures were brought to dryness by vacuum evaporation and subsequently resuspended in 100 µl methanol and centrifuged at 13,000 rpm for 15 min. Five µl of the supernatants were analyzed on LC-MS.

For isolation of the enzyme products, the reaction mixtures were scaled up to 10 ml, containing 50 mM Tris-HCl, pH 7.5, 10 mM $CaCl_2$, 2 mM DMAPP, 2-7.5 % (v/v) glycerol, 1-2 % (v/v) DMSO, 1 mM aromatic substrate and 0.4 mg·ml^{-1} purified recombinant protein, and incubated at 37 °C for 16 h. The reactions were terminated by addition of 10 ml methanol and brought to dryness by using a rotary evaporator at 37 °C. The residues were resuspended in 1 ml methanol, centrifuged at 13,000 rpm for 15 min, and purified on an HPLC device.

5.3. LC-ESI-HRMS analysis of the reaction mixtures

The treated enzyme reaction mixtures (5 µl) mentioned above were analyzed on an Agilent 1260 Infinity HPLC System (Böblingen, Germany) in combination with a photodiode array detector and a Bruker micrOTOF-Q III mass spectrometer. For separation, a Multospher 120 RP-18 column (250×2 mm, 5 µm, CS-Chromatographie Service Langerwehe, Germany) with a flow rate of 0.25 ml·min^{-1} was used. Water (solvent A) and MeCN (solvent B), both containing 0.1 % (v/v) formic acid, were used for a linear gradient of 5-100 % (v/v) solvent B in A in 40 min. Subsequently the column was washed with 100 % solvent B for 5 min and equilibrated with 5 % (v/v) solvent B for 10 min. The separations were monitored with the Bruker micrOTOF-Q III mass spectrometer using the positive-ion electrospray ionization (ESI). HPLC and MS data were processed by using Bruker Compass DataAnalysis Version 4.2 (Build 383.1) software.

5.4. Isolation of enzymatic products

Isolation of the enzyme products was performed on an Agilent HPLC series 1200. The separation was carried out on a MultoHigh Chiral AM-RP column (250×10 mm, 5 µm, CS-Chromatographie Service) with a flow rate of 1 ml·min^{-1} and different linear gradients of methanol in water.

5.5. NMR analysis

The isolated enzyme products were brought to dryness by using a rotary evaporator at 37 °C and dissolved in 0.7 ml CD_3OD. NMR spectra were recorded on a JEOL ECA-500 MHz spectrometer (JEOL Germany GmbH, Munich, Germany). The signal of CD_3OD at 3.31 ppm was used as internal reference for chemical shifts. Data processing was done by using MestReNova Version 6.0.2-5475 software.

Compound 21a: ^{1}H NMR (methanol-*d*$_4$, 500 MHz) δ = 7.55 (dt, *J* = 8.0, 0.9 Hz), 7.30 (dt, *J* = 8.2, 0.9 Hz), 7.12 (td, *J* = 8.2, 7.0, 0.9 Hz), 7.06 (s), 7.02 (td, *J* = 8.0, 7.0, 0.9 Hz), 5.35 (m), 4.70 (d, *J* = 6.8 Hz) Approx. 3.33 (t, *J* = 7.3 Hz, signal overlapping with those of solvent), 2.9 (t, *J* = 7.3 Hz), 2.89 (q, *J* = 7.4 Hz), 1.85 (s), 1.76 (s), 1.19 (t, *J* = 7.4 Hz); HR-ESI-MS: *m/z* = 321.1647, calcd. for $C_{17}H_{25}N_2O_2S$, $[M+H]^+$: 321.1631.

Compound 21b: ^{1}H NMR (methanol-*d*$_4$, 500 MHz) δ = 7.14 (d, *J* = 7.6 Hz), 7.05 (td, *J* = 7.8, 7.6, 1.1 Hz), 6.70 (td, *J* = 7.6, 7.6, 0.9 Hz), 6.60 (d, *J* = 7.6 Hz), 6.07 (dd, *J* = 17.4, 10.9 Hz), 5.38 (s), 5.12 (dd, *J* = 10.9, 1.3 Hz), 5.08 (dd, *J* =17.4, 1.3 Hz), 3.54 (dd, *J* = 10.0, 8.4 Hz), 3.10 (q, *J* = 7.4 Hz), 2.94 (ddd, *J* =11.5, 9.7, 5.3 Hz), 2.40 (ddd, *J* = 12.1, 11.9, 7.9 Hz), 2.07 (dd, *J* =12.3, 5.3 Hz), 1.29 (t, *J* = 7.4 Hz), 1.10 (s), 0.98 (s); HR-ESI-MS: *m/z* = 321.1642, calcd. for $C_{17}H_{25}N_2O_2S$, $[M+H]^+$: 321.1631.

Compound 30a: ^{1}H NMR (methanol-*d*$_4$, 500 MHz) δ = 7.46 (br d, *J* = 7.5 Hz), 7.37 (d, *J* = 8.6 Hz), 7.35 (br t, *J* = 7.5 Hz), 7.30 (br t, *J* = 7.5 Hz), 6.92 (d, *J* = 2.2 Hz), 6.81 (s), 6.73 (dd, *J* = 8.6, 2.2 Hz), 5.40 (m), 5.08 (s), 3.38 (d, *J* = 7.0 Hz), 1.76 (s), 1.74 (s); HR-ESI-MS: *m/z* = 292.1703, calcd. for $C_{20}H_{22}NO$, $[M+H]^+$: 292.1696.

Compound 30b: ^{1}H NMR (methanol-*d*$_4$, 500 MHz) δ = 7.48 (br d, *J* = 7.5 Hz), 7.37 (br t, *J* = 7.5 Hz), 7.29 (br t, *J* = 7.5 Hz), 7.24 (s), 7.04 (d, *J* = 3.2 Hz), 6.96 (s), 6.28 (dd, *J* = 3.2 Hz, 0.9), 5.35 (m), 5.09 (s), 3.39 (d, *J* = 7.5 Hz), 1.72 (s), 1.67 (s); HR-ESI-MS: *m/z* = 292.1704, calcd. for $C_{20}H_{21}NO$, $[M+H]^+$: 292.1696.

6. Acknowledgments

Shu-Ming Li and Peter Kolb are members of the LOEWE center for synthetic microbiology SYNMIKRO which provided the funding for Jakub Gunera’s PhD thesis work.

Peter Kolb thanks the German Research Foundation DFG for Emmy Noether fellowship K04095/1-1.

Shu-Ming Li also acknowledges the Deutsche Forschungsgemeinschaft for funding the Bruker microTOF-Q III mass spectrometer.

We thank Frank Balzer for help with making PrenDB available online and Christian Raab for data processing during the implementation.

References

1. Liu, A. H.; Liu, D. Q.; Liang, T. J.; Yu, X. Q.; Feng, M. T.; Yao, L. G.; Fang, Y.; Wang, B.; Feng, L. H.; Zhang, M. X.; and Mao, S. C. Caulerprenylols A and B, two rare antifungal prenylated para-xylenes from the green alga Caulerpa racemosa. *Bioorg. Med. Chem. Lett.* **2013,** *23,* 2491-2494.

2. Oya, A.; Tanaka, N.; Kusama, T.; Kim, S. Y.; Hayashi, S.; Kojoma, M.; Hishida, A.; Kawahara, N.; Sakai, K.; Gonoi, T.; and Kobayashi, J. Prenylated benzophenones from Triadenum japonicum. *J. Nat. Prod.* **2015,** *78* (258-264).

3. Sunassee, S. N.; Davies-Coleman, M. T. Cytotoxic and antioxidant marine prenylated quinones and hydroquinones. *Nat. Prod. Rep.* **2012,** *29,* 513-535.

4. Li, S.-M. Prenylated indole derivatives from fungi: structure diversity, biological activities, biosynthesis and chemoenzymatic synthesis. *Nat. Prod. Rep.* **2010,** *27,* 57-78.

5. Wollinsky, B.; Ludwig, L.; Hamacher, A.; Yu, X.; Kassack, M. U.; Li, S.-M. Prenylation at the indole ring leads to a significant increase of cytotoxicity of tryptophan-containing cyclic dipeptides. *Bioorg. Med. Chem. Lett.* **2012,** *22,* 3866-3869.

6. Botta, B.; Vitali, A.; Menendez, P.; Misiti, D.; Delle, M. G. Prenylated flavonoids: pharmacology and biotechnology. *Curr. Med. Chem.* **2005,** *12,* 717-739.

7. Winkelblech, J.; Fan, A.; Li, S.-M. Prenyltransferases as key enzymes in primary and secondary metabolism. *Appl. Microbiol. Biotechnol.* **2015,** *99,* 7379-7397.

8. Heide, L. Prenyl transfer to aromatic substrates: genetics and enzymology. *Curr. Opin. Chem. Biol.* **2009,** *13,* 171-179.

9. Mai, P.; Zocher, G.; Ludwig, L.; Stehle, T.; Li, S.-M. Actions of tryptophan prenyltransferases toward fumiquinazolines and their potential application for the generation of prenylated derivatives by combining chemical and chemoenzymatic syntheses. *Advanced Synthesis & Catalysis* **2016,** *358,* 1639–1653.

10. Fan, A.; Winkelblech, J.; Li, S.-M. Impacts and perspectives of prenyltransferases of the DMATS superfamily for use in biotechnology. *Appl. Microbiol. Biotechnol.* **2015,** *99,* 7399-7415.

11. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010,** *50* (5), 742-754.

12. Gunera, J.; Kolb, P. Fragment-based similarity searching with infinite color space. *Journal of Computational Chemistry* **2015,** *36* (21), 1597–1608.

13. Daylight Chemical Information System, Inc. http://www.daylight.com/dayhtml_tutorials/index.html (accessed March 16, 2016).

14. Weininger, D. SMILES, a chemical language and information system 1: Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988,** *28* (1), 31-36.

15. RDKit: Open source toolkit for chemoinformatics. http://www.rdkit.org (accessed March 16, 2016).

16. OpenEye Scientific Software, Inc., Sante Fe, NM, USA. http://www.eyesopen.com.

17. Irwin, J. J.; Shoichet, B. K. ZINC -- A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005,** *45,* 177-182.

18. McGann, M. FRED and HYBRID docking performance on standardized datasets. *Journal of Computer-Aided Molecular Design* **2012,** *26* (8), 897-906.

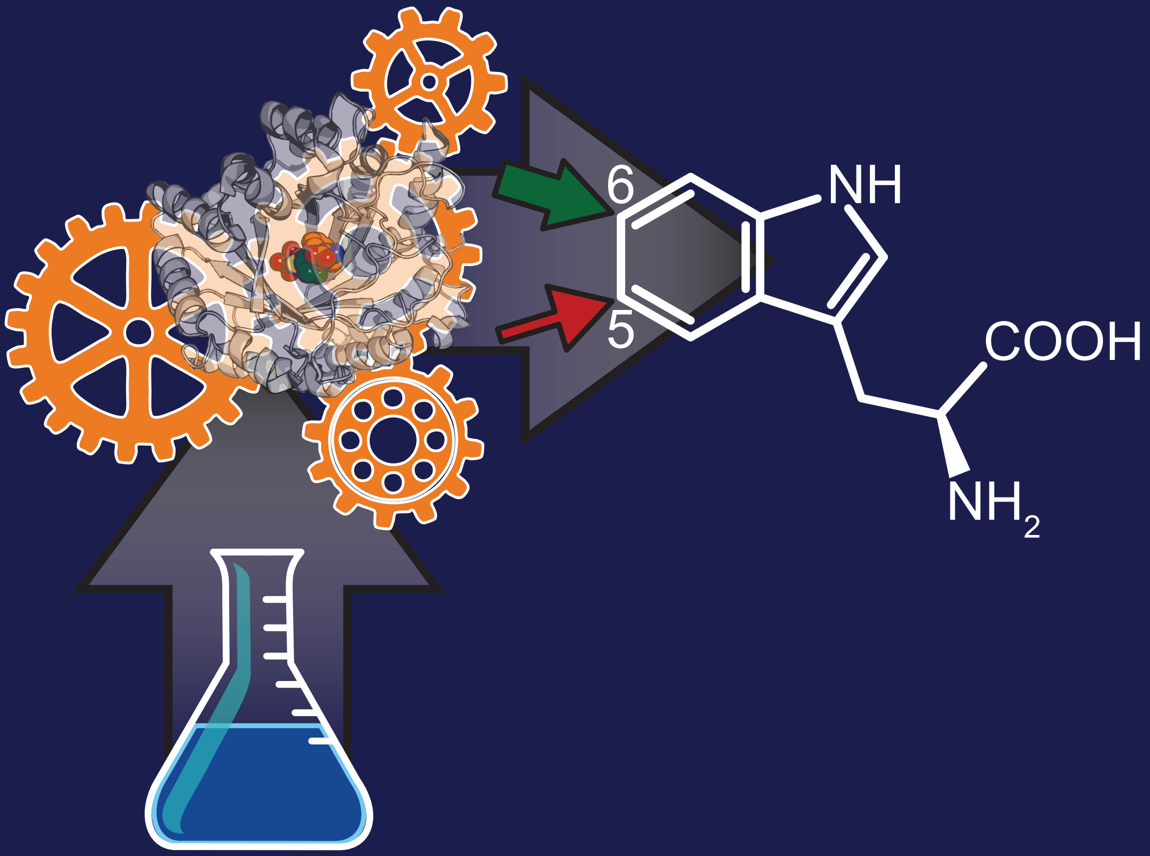
19. Woodside, A. B.; Huang, Z.; Poulter, C. D. Trisammonium geranyl diphosphate. *Org. Synth.* **1988,** *66,* 211-215.

20. Grundmann, A.; Li, S.-M. Overproduction, purification and characterization of FtmPT1, a brevianamide F prenyltransferase from Aspergillus fumigatus. *Microbiology* **2005,** *151,* 2199-2207.

21. Unsöld, I. A.; Li, S.-M. Overproduction, purification and characterization of FgaPT2, a dimethylallyltryptophan synthase from Aspergillus fumigatus. *Microbiology* **2005,** *151,* 1499-1505.

22. Yin, W.-B.; Ruan, H.-L.; Westrich, L.; Li, S.-M. CdpNPT, an N-prenyltransferase from Aspergillus fumigatus: overproduction, purification and biochemical characterisation. *ChemBioChem* **2007,** *8,* 1154-1161.

23. Schomburg, K.; Ehrlich, H.-C.; Stierand, K.; Rarey, M. From Structure Diagrams to Visual Chemical Patterns. *J. Chem. Inf. Model* **2010,** *50* (9), 1529-1535.



Advanced Synthesis and Catalysis **2015**, *357*, 975-986.

Part 3

Part 3 elaborates on the substrate selectivity and the regioselectivity of the prenyl transfer of five dimethylallyltryptophan synthases in the presence of unnatural prenyl donors. This investigation revealed a remarkable versatility of this enzyme family to not only accept different prenyl donors but also to transfer these moieties on different atoms within the acceptor molecule. With this study, we utilized a phalanx of *in silico* tools to elucidate the driving forces behind the observed regioselectivities and reaction yields. Thus, this work also contributes to a better understanding of prenyltransferases in general and to our *SAR-by-Enzyme* approach specifically.

The author list is the following (by contribution order): Winkelblech J, Liebhold M, Gunera J, Xie X, Kolb P and Li S.M. Julia designed and carried out the enzyme kinetic and enzyme activity assays. Mike conducted structure elucidation of the prenylated molecules via NMR and preformed cloning experiments on $\mathrm{DMATS}_{\mathrm{Sc}}$. I was responsible for the generation of a homology model of 5-DMATS, a dimethylallyltryptophan synthase. Furthermore, I conducted docking experiments in order to create initial geometries for Molecular Dynamics simulations and analyzed and interpreted the results thereof. Xiulan analyzed and interpreted the NMR data.

Tryptophan *C5*-, *C6*- and *C7*-prenylating enzymes displaying a preference for *C6* of the indole ring in the presence of unnatural dimethylallyl diphosphate analogs

Julia Winkelblech[1,2,‡], Mike Liebhold[1,‡], Jakub Gunera[2,3], Xiulan Xie[4], Peter Kolb[2,3] and Shu-Ming Li[1,2]

[1]Department of Pharmaceutical Biology and Biotechnology, Philipps-University, Marburg, Hesse, 35032, Germany

[2]Synmikro, LOEWE Centre for Synthetic Microbiology, Philipps-University, Marburg, Hesse, 35043, Germany

[3]Department of Pharmaceutical Chemistry, Philipps-University, Marburg, Hesse, 35032, Germany

[4]Department of Chemistry, Philipps-University, Marburg, Hesse, 35032, Germany

[‡]These authors contributed equally to this study.

Keywords:

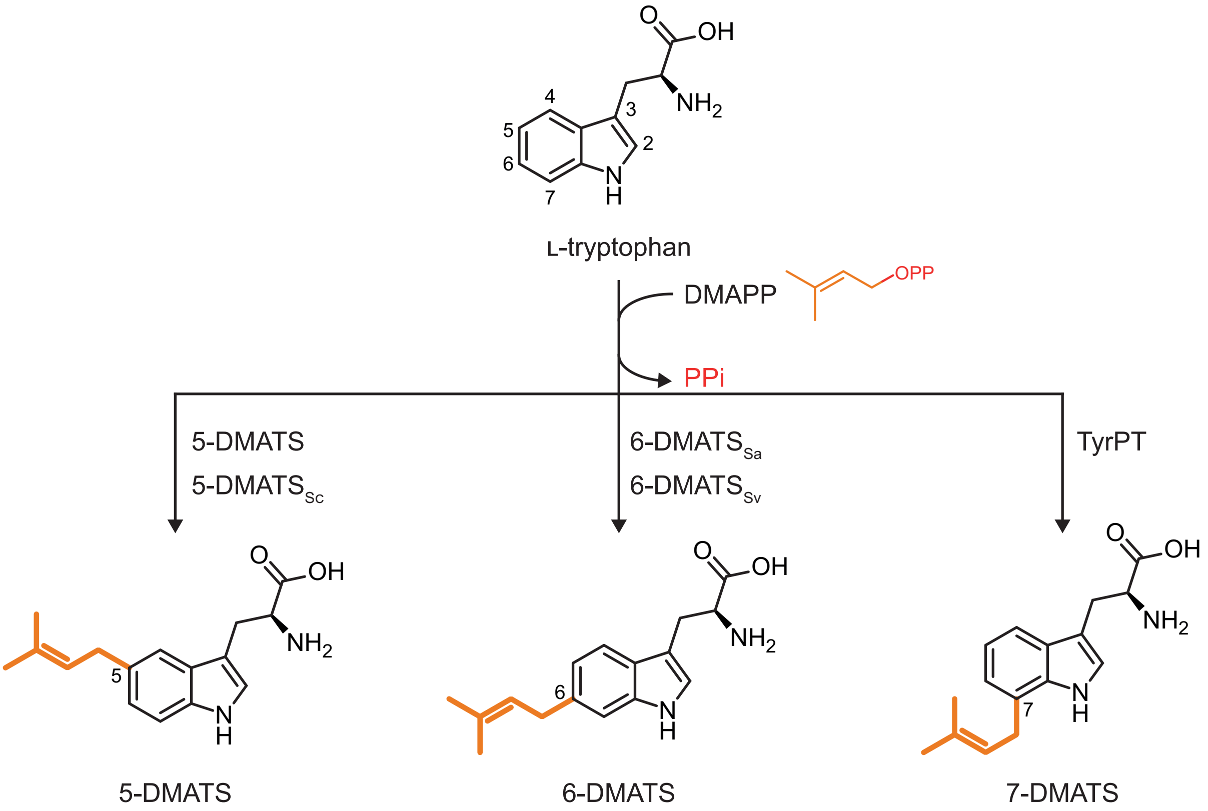
dimethylallyltryptophan synthase, enzyme catalysis, Friedel–Crafts alkylation, prenyltransferase, regioselectivity

Abstract

The behavior of four dimethylallyltryptophan synthases (DMATSs) (5-DMATS and 5-DMATS$_{Sc}$ as tryptophan C5-prenyltransferases, and 6-DMATS$_{Sa}$ and 6-DMATS$_{Sv}$ as C6-prenyltransferases) and one L*-tyrosine prenyltransferase with a tryptophan C7-prenyltransferase activity was investigated in the presence of two unnatural alkyl donors (methylallyl and 2-pentenyl pyrophosphate) and one benzyl donor (benzyl pyrophosphate). Detailed biochemical investigations revealed the acceptance of these dimethylallyl pyrophosphate (DMAPP) analogs by all tested enzymes with different relative activities. Enzyme products with the allyl or benzyl moiety attached to different positions were identified in the reaction mixtures, whereby C6-alkylated or benzylated* L*-tryptophan was found as one of the main products. This observation demonstrates a preference of the five prenyltransferases toward C6 of the indole ring in the presence of unnatural DMAPP derivatives. Molecular dynamics simulation experiments with a homologous model of 5-DMATS explained well its reactions with methylallyl and 2-pentenyl pyrophosphate. Furthermore, this study expands significantly the potential usage of tryptophan prenylating enzymes as biocatalysts for Friedel–Crafts alkylation.*

1. Introduction

Secondary metabolites with biological activities represent an important source for medicinal research and drug development. (1,2) They are widely distributed in nature, especially in plants and microorganisms. (1,2,3) Among microorganisms, fungi of Ascomycetes and bacteria of Actinomycetes are important producers of biologically active compounds. (3,4) Due to significant progress in genome sequencing and genome mining, a



Scheme 1: Regiospecific prenylation of tryptophan by the five prenyltransferases used in this study in the presence of their natural prenyl donor DMAPP. Origin of the enzymes: 6-DMATS_{Sa} from *Streptomyces ambofaciens*, 6-DMATS_{Sv} from *Streptomyces violaceusniger*, 5-DMATS from *Aspergillus clavatus*, 5-DMATS_{Sc} from *Streptomyces coelicolor* and TyrPT from *Aspergillus niger*.

number of gene clusters involved in the biosynthesis of such metabolites have been identified in recent years. (4,5,6,7,8) A large group of natural products comprises the prenylated aromatic substances derived from prenyl diphosphate and an aromatic scaffold from different pathways. (9,10) Prenyltransferases catalyze the linkage of these two residues and play an important role in the structural diversity of these compounds. Indole prenyltransferases belong to the ***di****m*ethyl***a****llyl**t*ryptophan ***s****ynthase* (DMATS) superfamily, which catalyze the underlying prenylation reaction of indole derivatives in nature, and represent one of the most investigated class of prenyltransferases. (11) In the presence of the natural prenyl donor dimethylallyl pyrophosphate (DMAPP), most members of this superfamily usually show remarkable flexibility toward their aromatic substrates, but high regioselectivity of the prenylation position on the indole ring. (12,13,14,15) These characteristics were observed for fungal tryptophan prenyltransferases, *e.g.*, FgaPT2, 5-DMATS and 7-DMATS from different *Aspergillus* spp., which catalyze tryptophan *C4*-, *C5*- and *C7*-prenylations, respectively (**Scheme 1**). Two bacterial enzymes, SCO7467 from *Streptomyces coelicolor* A3(2) and IptA from *Streptomyces* sp. SN-593, are tryptophan *C5*- and *C6*-prenyltransferases, respectively. (12,16) IptA is involved in the biosynthesis of 6-dimethylallylindole-3-carbaldehyde. (12,17)

Recently, two further 6-DMATS enzymes, 6-DMATS_{Sa} (SAML0654) from *Streptomyces ambofaciens* (*S. ambofaciens*) ATCC238 and 6-DMATS_{Sv} (Strvi8510) from *Streptomyces violaceusniger (S. violaceusniger)* Tü4113 were identified and characterized biochemically. (18) These two 6-DMATS enzymes showed high flexibility toward their prenyl donor and acceptor. In contrast to other indole prenyltransferases, both DMAPP and geranyl pyrophosphate (GPP) were used by both enzymes. (18) Consequently, this flexibility makes them interesting candidates for further investigations on the acceptance of unnatural alkyl or benzyl donors.

Biochemical investigations on the tryptophan prenyltransferases FgaPT2 and 5-DMATS with methylallyl (MAPP) and 2-pentenyl pyrophosphate (2-pentenyl-PP or EAPP for ethylallyl pyrophosphate) showed that these enzymes also accepted such unnatural alkyl donors. The alkylation positions were shifted partially or completely to the neighboring position. (19) The tryptophan *C4*-prenyltransferase FgaPT2 even accepted benzyl pyrophosphate (benzyl-PP or BENZYLPP) as substrate and catalyzed the regiospecific benzylation of L-tryptophan at position *C5*. (20) Our previous data on the reactions of tryptophan prenyltransferases with unnatural alkyl and benzyl donors were limited to enzymes which catalyzed the transfer reactions of the dimethylallyl moiety onto position *C4* and *C5* of the indole ring. (19,20) In a previous study, the behavior of the 7-DMATS from *A. fumigatus* could not be investigated in detail, due to its low activity in the presence of unnatural DMAPP analogs. (19)

Fortunately, the recently identified L-tyrosine prenyltransferase TyrPT from *Aspergillus niger* showed a remarkable tryptophan *C7*-prenyltransferase activity (21) and can be considered as a tryptophan *C7*-prenylating enzyme in this study. As shown in **Scheme 1**, the five enzymes 5-DMATS (13), 5-DMATS$_{Sc}$ (17), 6-DMATS$_{Sa}$ (18), 6-DMATS$_{Sv}$ (18) and TyrPT (21) used in this study share the same substrates (tryptophan and DMAPP), but catalyzed regiospecific prenylations at different positions of the indole ring. After having the availability of the two tryptophan *C6*-prenyltransferases 6-DMATS$_{Sa}$ and 6-DMATS$_{Sv}$ as well as of TyrPT in our laboratory, we initiated a study to prove their behavior toward MAPP, EAPP and BENZYLPP.

2. Results

2.1. *C6*-alkylated/benzylated derivates as unique enzyme products of the two *C6*-prenyltransferases

The purified recombinant proteins 6-DMATS$_{Sa}$ and 6-DMATS$_{Sv}$ were firstly incubated with L-tryptophan in the presence of one of the three unnatural DMAPP analogs MAPP (**I**), EAPP (**II**), and BENZYLPP (**III**). HPLC analysis of the enzyme assays showed clear product formation in all of these reaction mixtures, with the highest conversions of 91.3±0.07 % and 89.3±0.6 % observed in the presence of EAPP for 6-DMATS$_{Sa}$ and 6-DMATS$_{Sv}$, respectively (**Figure 1** see the Supporting Information, **Table S1**). Lower conversion yields of 51.1±0.5 % and 37.6±0.3 % were observed in the incubation mixtures with MAPP, and 13.9±0.3 % and 8.2±0.3 % with BENZYLPP (**Figure 1**, **Scheme 2**, see the Supporting Information, **Table S1**). To determine the alkylation position, enzyme assays were prepared on a large scale. The enzyme products **Ia** – **IIIa** were isolated from both assays of 6-DMATS$_{Sa}$ and 6-DMATS$_{Sv}$ on HPLC and their structures were elucidated by MS and NMR analyses. For better understanding, we named the products by a combination of **I** (product from MAPP), **II** (EAPP) or **III** (BENZYLPP) with **a** (regular alkyl or benzyl at *C6*), **b1** (regular alkyl at *C7*), **b2** (reverse alkyl at *C7*), **b** (benzyl at *C7*) or **c** (regular alkyl or benzyl at *C5*). MS data confirmed the monoalkylation or benzylation of the isolated products. ^{1}H NMR analysis (for structural elucidation see the Supporting Information) proved the regular attachment of the alkyl or benzyl residue onto position *C6* of the indole ring in all of these cases (see the Supporting Information, **Figures S9** and **S10**). This conclusion was drawn by comparison of the coupling patterns of the signals for aromatic protons with those of the published data for *C6*-alkylated L-tryptophan. (18,19,20) In the presence of the natural prenyl donors DMAPP or GPP, 6-DMATS$_{Sa}$ and 6-DMATS$_{Sv}$ also catalyze a *C6*-prenylation. (18) Therefore, the alkylation position for both enzymes was proven to be independent of the used alkyl or benzyl donor.

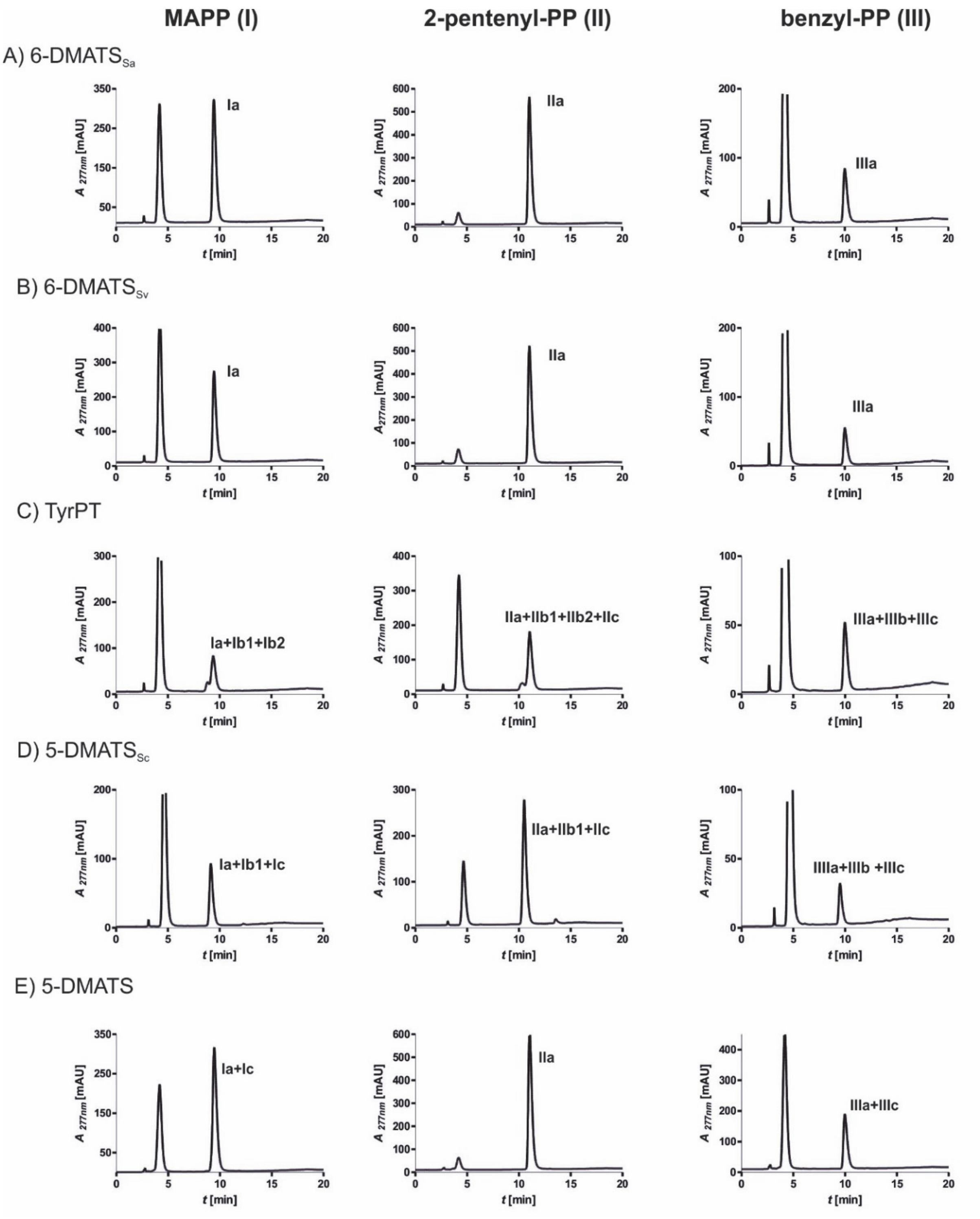


Figure 1: HPLC analysis of the reaction mixtures of L-tryptophan with unnatural DMAPP analogs.

2.2. *C6*-alkylated/benzylated derivates were main products of TyrPT reactions with DMAPP analogs

Taking the data on 6-DMATS$_{Sa}$ and 6-DMATS$_{Sv}$ with the previous published results on FgaPT2 and 5-DMATS (19,20,22) together, we have shown the behavior of tryptophan *C4*-, *C5*- and *C6*-prenyltransferases toward unnatural DMAPP analogs. It would be interesting to complete this series with *C7*-prenylating enzymes. A previous study showed that the tryptophan *C7*-prenyltransferase 7-DMATS from *A. fumigatus* (23) accepted very poorly MAPP and EAPP. (19) Recently, CAK41583 from *A. niger* was identified as a tyrosine prenyltransferase

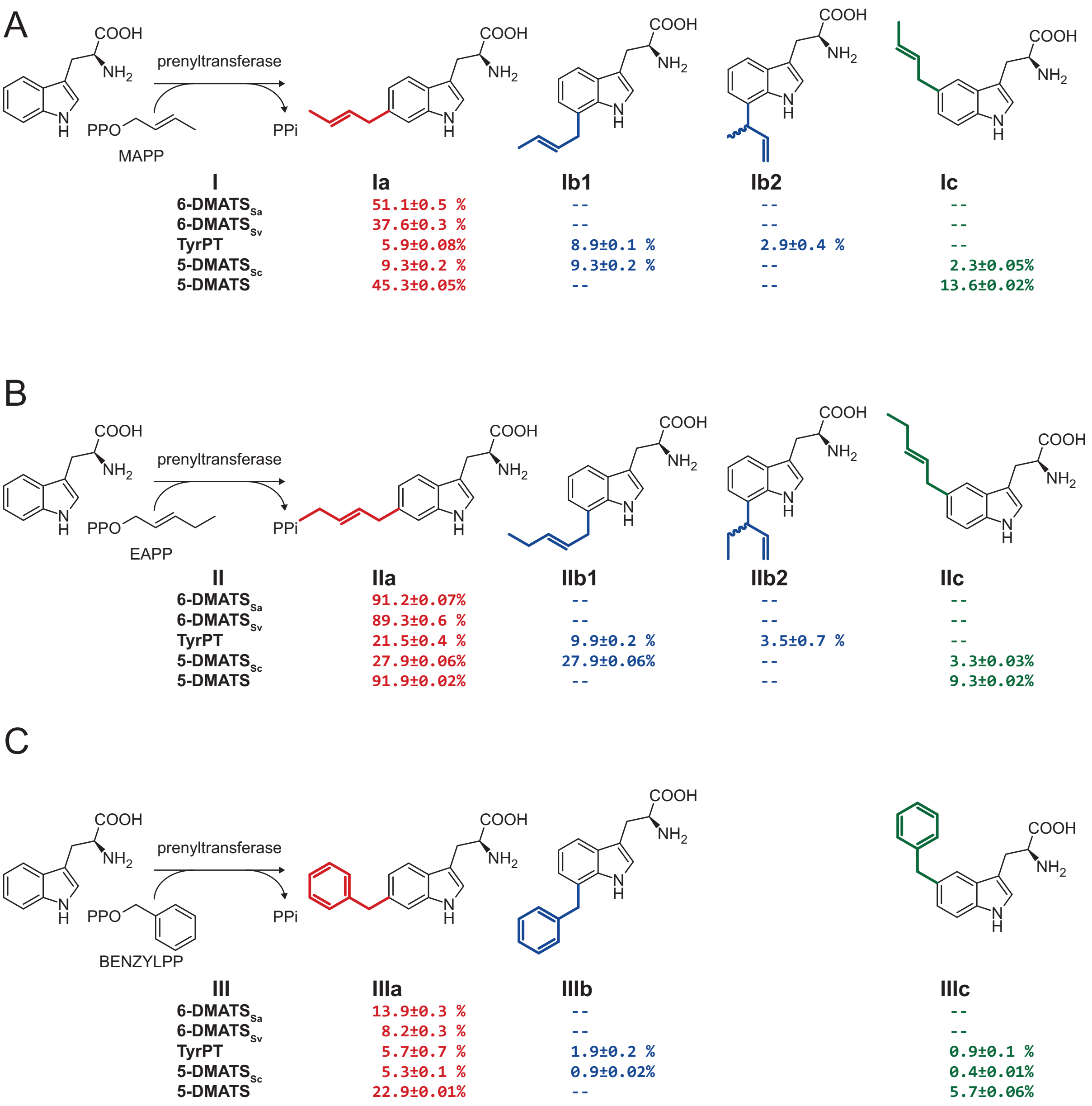
Table 1: Kinetic parameters of the tested prenyltransferases toward DMAPP and analogs thereof (MAPP, EAPP and BENZYLPP).

	DMAPP		EAPP		MAPP		BENZYLPP	
	K_M [mM]	k_{cat} [min^{-1}]	K_M [mM]	k_{cat} [min^{-1}]	K_M [mM]	k_{cat} [min^{-1}]	K_M [mM]	k_{cat} [min^{-1}]
6-DMATS$_{Sa}$	0.095±0.011^a	37.8±4.1^a	0.011±0.000094	0.18±0.0043	0.025±0.0019	0.066±0,00088	0.036±0.00049	0.074±0.00065
6-DMATS$_{Sv}$	0.025±0.0005	9.9±0.13	0.049±0.0021	0.24±0.0018	0.040±0.0018	0.064±0.00036	0.081±0.0021	0.047±0.0027
5-DMATS$_{Sc}$	0.05±0.002^b	24±6^b	0.028±0.0011	0.082±0.0035	0.054±0.0004	0.029±0.00014	0.11±0.019	0.017±0.0022
TyrPT	0.39±0.018^c	0.22±0.023^c	0.033±0.00078	0.056±0.0003	0.026±0.00024	0.067±0.00077	0.02±0.00042	0.029±0.00058
5-DMATS	0.076^d	78^d	0.13±0.021^e	1.38±0.12^e	0.04±0.003^e	0.3±0.024^e	0.12±0.0014	0.092±0.00054

[a-e]Data were adapted from previous publications for 6-DMATS$_{Sa}$ (18), 5-DMATS$_{Sc}$ (17), TyrPT (21), 5-DMATS (13,19) Data for DMAPP analogs are mean values with difference range obtained from two independent measurements.

(TyrPT), catalyzing an *O*-prenylation at the phenolic hydroxy group of L-tyrosine. (21) As in the case of SirD from *Leptosphaeria maculans*, (24) TyrPT also catalyzed the transfer reaction of a dimethylallyl moiety from DMAPP to *C7* of L-tryptophan and several derivatives thereof. (21) The broad substrate specificity of TyrPT led us to test its activity for DMAPP analogs in the presence of L-tryptophan. In analogy to 6-DMATS$_{Sa}$ and 6-DMATS$_{Sv}$, TyrPT was incubated with L-tryptophan in the presence of MAPP, EAPP and BENZYLPP. Product formation was detected in all three incubation mixtures (**Figure 1**). However, the observed enzyme activities were much lower than those of the two 6-DMATS enzymes. Total product yields of 38.3±0.6, 17.7±0.2 and 8.5±1.0 % were calculated for EAPP, MAPP and BENZYLPP, respectively (**Figure 1**, **Scheme 2**, see the Supporting Information, **Table S1**). This is justified by the fact that L-tyrosine, but not L-tryptophan is the best accepted aromatic substrate by TyrPT, also in the presence of DMAPP. Interestingly, the ratio of the relative activities toward the three DMAPP analogs was similar to those of the two 6-DMATS enzymes. In contrast to the unique *C7*-prenylation of L-tryptophan by TyrPT in the presence of DMAPP, interpretation of the individual peaks of the ^{1}H NMR spectra indicated the presence of more than one product each in the incubation mixtures with DMAPP analogs. Optimization of the HPLC conditions and the application of a Chiralpak Zwix (+) column (see the Supporting Information, **Figure S 1**) allowed a partial separation of these product mixtures. Although the compounds to be separated differ from each other by alkylation positions rather than by stereochemistry, they showed different behavior on the Chiralpak Zwix (+) column. It seems that the indole derivatives had different interactions with the column material.

HPLC analysis of the incubation mixture with EAPP and interpretation of the NMR data led to the identification of three substances with a regular alkyl moiety attached to *C6* (**IIa**), *C7* (**IIb1**) and *C5* (**IIc**), respectively. Product yields of 21.5±0.4, 9.9±0.2 and 3.3±0.03 % were calculated for these products (**Scheme 2**, see also the Supporting Information). In addition, a



	Ia	Ib1	Ib2	Ic
I				
6-DMATS$_{Sa}$	51.1±0.5 %	--	--	--
6-DMATS$_{Sv}$	37.6±0.3 %	--	--	--
TyrPT	5.9±0.08%	8.9±0.1 %	2.9±0.4 %	--
5-DMATS$_{Sc}$	9.3±0.2 %	9.3±0.2 %	--	2.3±0.05%
5-DMATS	45.3±0.05%	--	--	13.6±0.02%

	IIa	IIb1	IIb2	IIc
II				
6-DMATS$_{Sa}$	91.2±0.07%	--	--	--
6-DMATS$_{Sv}$	89.3±0.6 %	--	--	--
TyrPT	21.5±0.4 %	9.9±0.2 %	3.5±0.7 %	--
5-DMATS$_{Sc}$	27.9±0.06%	27.9±0.06%	--	3.3±0.03%
5-DMATS	91.9±0.02%	--	--	9.3±0.02%

	IIIa	IIIb	IIIc
III			
6-DMATS$_{Sa}$	13.9±0.3 %	--	--
6-DMATS$_{Sv}$	8.2±0.3 %	--	--
TyrPT	5.7±0.7 %	1.9±0.2 %	0.9±0.1 %
5-DMATS$_{Sc}$	5.3±0.1 %	0.9±0.02%	0.4±0.01%
5-DMATS	22.9±0.01%	--	5.7±0.06%

Scheme 2: Alkylation/benzylation of L-tryptophan catalyzed by *C5*-, *C6*- and *C7*-prenylating enzymes in the presences of three DMAPP analogs: **A** methylallyl pyrophosphate (MAPP); **B** 2-pentenyl pyrophosphate (EAPP); **C** benzyl diphosphate (BENZYLPP). --: product yields < 0.3 %. The mean of the total conversion yields was measured in duplicate by HPLC and the percentages for different products were calculated by using corresponding NMR data.

reversely *C7*-alkylated L-tryptophan (**IIb2**) was isolated with a product yield of 3.5±0.7 % (**Scheme 2**, see the Supporting Information, **Figure S21**). With MAPP as alkyl donor, regularly *C6*- (**Ia**) and *C7*-alkylated (**Ib1**) as well as reversely *C7*-alkylated derivatives (**Ib2**) were identified by interpretation of their NMR spectra (see the Supporting Information). Product yields of 5.9±0.08, 8.9±0.1 and 2.9±0.4 % were calculated for **Ia**, **Ib1** and **Ib2**, respectively. Regularly *C6*-, *C7*- and *C5*-benzylated products (**IIIa**, **IIIb**, **IIIc**) with product yields of 5.7±0.7, 1.9±0.2 and 0.9±0.1 % were identified in the reaction mixture of L-tryptophan with BENZYLPP. These results demonstrated clearly that *C6*-alkylated or benzylated derivatives were unique or one of two predominant products of TyrPT reactions in the presence of the unnatural donors (**Scheme 2**) and differed clearly from that of L-tryptophan with DMAPP. (21)

2.3. Comparison of bacterial and fungal alkylation/benzylation reactions by investigations on 5-DMATS and 5-DMATS$_{Sc}$

As described above, the two 6-DMATS enzymes from bacteria catalyzed the regiospecific alkylation and only one product with the same position, *i.e.*, *C6*, was identified, independent of DMAPP, GPP (18) or the DMAPP analogs MAPP, EAPP or BENZYLPP. In comparison, the fungal prenyltransferases FgaPT2 and 5-DMATS catalyzed the regiospecific *C4*- and *C5*-prenylation in the presence of DMAPP, respectively. (13,14) But in the presence of the unnatural DMAPP analogs, the regioselectivity was partially or completely shifted. (19,20) In the presence of L-tryptophan, the fungal L-tyrosine *O*-prenyltransferase TyrPT also accepted DMAPP analogs as substrates. In the case of EAPP and BENZYLPP, *C6*-alkylated or benzylated L-tryptophan was the predominant product. In the presence of MAPP, the *C6*-alkylated derivative was one of the two dominant products. It seems that in the presence of DMAPP analogs, *C6* is the preferable alkylation position for enzymes which usually catalyzed the prenylation of L-tryptophan at *C5* (like 5-DMATS), *C6* (6-DMATS enzymes) and *C7* (TyrPT).

These results pose an important question on the possible reason for the decreasing regioselectivity of 5-DMATS and TyrPT. One plausible explanation could be the orientation of the DMAPP analogs in the binding sites of the enzymes, which make *C6* to be the preferable alkylation position. However, it cannot be excluded that the observed regiospecific alkylation or benzylation of L-tryptophan at the same position by 6-DMATS$_{Sa}$ and 6-DMATS$_{Sv}$ in the presence of DMAPP, GPP and DMAPP analogs is based on their bacterial origin. It could be speculated that bacterial prenyltransferases retain their regioselectivities independent of the used donors, while fungal enzymes exhibit relaxed selectivity in the presence of different alkyl donors. 5-DMATS and TyrPT are fungal enzymes and therefore showed different behavior regarding regioselectivity compared with 6-DMATS enzymes. The latter hypothesis would also be supported by the fact that FgaPT2 catalyzed tryptophan alkylation and benzylation in the presence of these unnatural alkyl and benzyl donors with partial or complete shift of the attachment positions. (19,20)

To clarify the possible reason for this difference, we investigated the regioselectivity of the tryptophan *C5*-prenyltransferase SCO7467 (5-DMATS$_{Sc}$) from the bacterium *Streptomyces coelicolor* A3(2) in the biosynthesis of 5-dimethylallylindole-3-acetonitrile. (16,17)

SCO7467 was overproduced in *E. coli* as reported by Ozaki (17), purified and investigated in the presence of MAPP, EAPP and BENZYLPP. For comparison, the behavior of the fungal 5-DMATS from *A. clavatus* (13), toward MAPP and EAPP (19) was reproduced in this study. In addition, this enzyme was assayed with BENZYLPP in the presence of L-tryptophan.

The previously reported data for 5-DMATS (19) were reproduced in this study by identification of *C5*- and *C6*-alkylated products with MAPP, with products yields of 13.6±0.02 and 45.3±0.05 %, respectively. In the presence of EAPP, the alkylation position was completely shifted from *C5* to *C6*. Similar to those of MAPP, *C5*- and *C6*-benzylated products with yields of 5.7±0.01 and 22.9±0.06 % were detected in the assay with BENZYLPP (**Figure 1**, **Scheme 2**). Again, *C6*-alkylated or benzylated L-tryptophan represented the predominant product.

HPLC analysis clearly revealed product formation in the reaction mixtures of L-tryptophan with the recombinant 5-DMATS$_{Sc}$ in the presence of all three DMAPP analogs (**Figure 1**). HR-MS data confirmed the attachment of one alkyl or benzyl residue on the substrate for all of the obtained products (see the Experimental Section). Structure elucidation by NMR indicated that the isolated product peaks consisted of more than one substance. *C6*-, *C7*- and *C5*-alkylated derivatives were identified with ratios of 3:3:1 for EAPP and of 4:4:1 for MAPP. By using a Chiralpak Zwix (+) column, the *C7*-alkylated products were purified from these mixtures (**Scheme 2**, see the Supporting Information, **Figure S1**). With EAPP as alkyl donor, a product yield of 9.3±0.02 % was calculated for *C5*- (**IIc**) and 27.9±0.06 % each for *C6*- (**IIa**) and *C7*-alkylated (**IIb1**) L-tryptophan (**Scheme 2**). In the case of MAPP, product yields

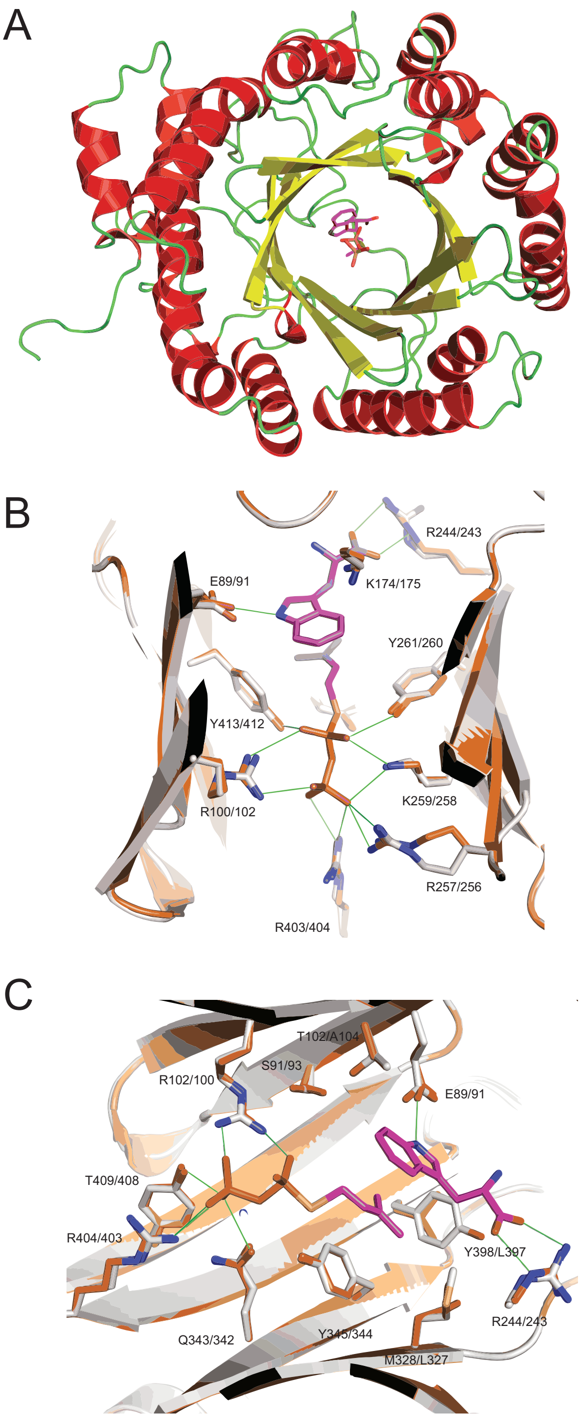


Figure 2: Homology model of 5-DMATS (**A**). α-helices are colored in red, β-sheets in yellow and turns and loop in green, respectively. The ABBA motif of dimethylallyltryptophan synthases is reproduced in the model, the Cα-RMSD between model and template being 0.1 Å. Active site residues of the model (orange) and template (white) are shown (**B** and **C**). The corresponding amino acids are labelled as pairs (FgaPT2/5-DMATS).

of 9.3±0.2, 9.2±0.2 and 2.3±0.05 % were determined for **Ia**, **Ib1** and **Ic**, respectively. Inspection of the NMR spectra of the products obtained with BENZYLPP revealed the presence of 6-benzyl-L-tryptophan (**IIIa**) with a product yield of 5.3±0.1 % and 7-benzyl-L-tryptophan (**IIIb**) of 0.9±0.02 % (see the Supporting Information, **Figures S11** and **S22**). In addition, signals of a *C5*-benzylated L-tryptophan (**IIIc**) with a product yield of 0.4±0.009 % could also be observed (together with **IIIa** as a mixture, see the Supporting Information, **Figure S2**).

The results obtained with the bacterial 5-DMATS$_{Sc}$ were distinguishable not only from those with the fungal 5-DMATS, but also from those of the two bacterial 6-DMATS enzymes. Formation of three different alkylated or benzylated products by 5-DMATS$_{Sc}$ in all of the three incubations disproved the bacterial origin of the observed high regiospecificity for the two 6-DMATS enzymes. These results confirmed the preference of the enzymes investigated in this study for *C6* of the indole ring in the presence of the three unnatural DMAPP analogs.

2.4. Kinetic parameters

Determination of the kinetic parameters of the enzymes with the DMAPP analogs indicated that the observed reactions were consistent with Michaelis–Menten kinetics (**Table 1**). K_M values in the range of 0.011 to 0.13 mM proved their relatively high affinity toward the tested DMAPP analogs. In contrast, the turnover numbers of the reactions with these DMAPP analogs were much lower than those with DMAPP. As observed in **Figure 1** and given in **Scheme 2**, EAPP was accepted in most cases as the best unnatural alkyl donor. This was also confirmed by the kinetic parameters with an exception for TyrPT. Here the efficiencies toward MAPP and EAPP are almost identical, although higher relative activities toward EAPP were observed. The unnatural donor MAPP was also well accepted but to a lesser degree. BENZYLPP is a poor substrate for all enzymes, as verified by kinetic parameters.

2.5. Homology modelling of 5-DMATS

To get insights into the reduced regioselectivity of the tested enzymes in the presence of DMAPP analogs and to assess how the protein might be able to discriminate between the different

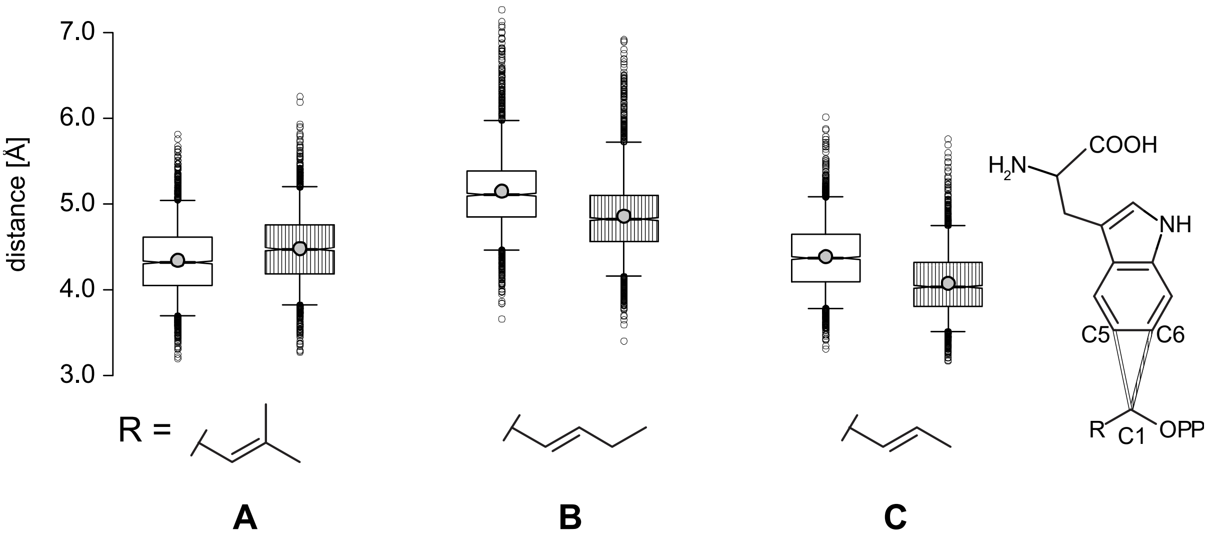


Figure 3: For each donor molecule the distances between *C1* atom of the donor and *C5*, and *C6* atom of tryptophan were measured over 2,500 generated coordinates. Distance distributions are shown as box plots: Grey circles represent mean values, white circles measurements outside the 95th percentile. Boxes span 50 % of the measurements, whiskers 95 %. White boxes correspond to distances between *C1* and *C5* atom, while shaded boxes show distances between *C1* and *C6* atom, respectively.

analogs, we homology modelled 5-DMATS. Four enzymes from the DMATS superfamily, FgaPT2 (25), FtmPT1 (26), CdpNPT (27) and AnaPT (28) could principally serve as templates. As expected, the structure of the tryptophan *C4*-prenyltransferase FgaPT2 is the most suitable for this purpose, owing to the sequence identity of 52 % with the target. As shown in **Figure 2A** and **Figure S26** in the Supporting Information, our model of 5-DMATS consists of five *αββα* units, being similar to those of the known structures of the DMATS enzymes. Due to the low homology of only about 26 % or less on the amino acid level to proteins with known structures, no model with a sufficient level of detail for the approaches used in this study could be obtained for 5-DMATS$_{Sc}$, 6-DMATS$_{Sa}$, 6-DMATS$_{Sv}$, or TyrPT.

2.6. Docking experiments with DMAPP and analogs

Initial docking experiments led to acceptor and donor poses consistent with the interactions observed for the respective molecules in the template X-ray structure. In particular, contacts with the conserved basic residues interacting with the pyrophosphate tail of the donor molecules are preserved. Yet, this static picture of protein acceptor – donor interactions did not allow us to formulate a hypothesis that was consistent with the experimental findings. Thus, we carried out molecular dynamics (MD) calculations to assess how the interactions might change over time.

These MD studies (Supporting Information, **Figure S27**) showed that DMAPP resides in the cavity with a mean distance of 4.32 Å between its *C1* and the *C5* of the indole ring (**Figure 3A**). In contrast, the average distance between *C1* and *C6* on L-tryptophan is significantly larger, thus providing a possible explanation for the formation of solely *C5*-prenylated tryptophan in the presence of DMAPP. In comparison, *C1* of EAPP is predominantly close to *C6* of the indole ring with a distance of 4.82 Å (compared to 5.11 Å between *C1* and *C5*; **Figure 3B**), so that an exclusive *C6*-alkylation is plausible. As shown in **Figure 3C**, *C1* of MAPP is located at a shorter distance to *C6* of the indole ring, which is also consistent with the formation of the predominantly *C6*-alkylated derivative for this donor. All these simulations were remarkably stable over the simulation time, as evidenced by the RMSD and RMSF plots in **Figure S28** and **S29** of the Supporting Information and the overlay of the starting structure and the final snapshot (Supporting Information, **Figure S27**). In contrast, the MD simulations with BENZYLPP became unstable shortly after the start of the unrestrained equilibration step (despite several repetitions), with an unusual edge-to-face orientation of benzyl pyrophosphate with respect to L-tryptophan (Supporting Information, **Figure S27**). We were thus unable to use these data within the present study. We speculate that the reason for this behavior of the simulations could be that the binding sites for the prenyl donors in the structures of the DMATS enzymes were determined with linear DMAPP analogs, which have significantly different sizes and electron densities than benzyl pyrophosphate.

It also seems intuitive that the donor-acceptor distance is a major factor determining the regioselectivity of alkylations: As evidenced by the experiments, the reactivities of both *C5* and *C6* can be considered as equal. Thus, the preference for prenylation at a certain position can be a direct effect of the number of times donor and acceptor come so close to each other that an activated complex can be formed.

3. Discussion and conclusions

In conclusion, all the tested enzymes used methylallyl, 2-pentenyl and benzyl pyrophosphate as substrates and catalyzed Friedel–Crafts alkylation or benzylation reactions on the indole ring. The observed reactions differ from each other in relative activities and regioselectivity of the attached position. One to four alkylated or benzylated derivatives have been identified as enzyme products (**Scheme 2**). From **Scheme 2**, it is obvious that in the presence of unnatural DMAPP analogs, *C6* of tryptophan was the preferable alkylation and benzylation position for tryptophan *C5*-, *C6*- and *C7*-prenylating enzymes. *C6*-Alkylated or benzylated derivatives were identified in all the reaction mixtures. It was found as a unique product in the cases of the both 6-DMATS enzymes with all of the three DMAPP analogs or as one of two main products in the reaction mixtures of 5-DMATS$_{Sc}$ with MAPP and EAPP. Such derivatives were predominant products in all other reaction mixtures. From **Scheme 2**, it is also clear that the tryptophan *C5*-prenyltransferases 5-DMATS and 5-DMATS$_{Sc}$ as well as TyrPT with a tryptophan *C7*-prenyltransferase activity also produced *C5*- or/and *C7*-alkylated or benzylated derivatives, indicating a shift of the alkylation or benzylation position from *C5* to *C7* and *vice versa*. By using the program MODELLER, a structural model was constructed for 5-DMATS from *A. clavatus* and used for docking and MD studies with DMAPP, MAPP and EAPP, leading to a distance-based explanation of their observed reaction preferences. Unfortunately, the MD simulations with benzyl pyrophosphate became unstable. It seems that the available structure information is still too limited for a universal interpretation or prediction of all possible enzyme reactions. Therefore, it will be interesting to have more protein structures elucidated in the near future, most importantly also as complexes with different acceptors and donors including unnatural DMAPP analogs.

4. Materials and Methods

4.1. Chemicals

Synthesis of methylallyl-PP (MAPP), 2-pentenyl-PP and benzyl-PP was carried out as described previously. (19,29) L-tryptophan was purchased from Roth (Karlsruhe, Germany).

4.2. Overproduction and purification of the recombinant proteins

Gene expression and subsequent protein purification of the recombinant 6-DMATS$_{Sa}$-His6, His8-6-DMATS$_{Sv}$, His6-TyrPT and 5-DMATS-His6 were carried out as described previously. (13,18,21)

4.3. Cloning and expression of 5-DMATS$_{Sc}$ (SCO7467)

PCR amplification of SCO7467 from Streptomyces coelicolor A3(2) was carried out as described by Ozaki *et al.* (30) The expression vector pHis8 containing the coding sequence was termed pML10. *E. coli* BL21 [DE3] cells harboring pML10 were cultivated in 1 L liquid lysogeny broth (LB) medium supplemented with kanamycin (50 µg ml^{-1}) till an absorption at 600 nm of 0.6. For induction of gene expression, IPTG was added to a final concentration of 0.5 mM. After further incubation at 30°C and 220 rpm for 6 h, the recombinant protein was purified as routinely on Ni-NTA agarose.

4.4. Enzyme assays for determination of the activities and kinetic parameters

The reaction mixtures (100 µl) for determination of the enzyme activities contained 1 mM L-tryptophan, 5 mM $CaCl_2$, 2 mM alkyl diphosphate (DMAPP, MAPP (**I**), EAPP (**II**)) or BENZYLPP (**III**), 1.0-1.5% (v/v) glycerol, 50 mM Tris-HCl (pH 7.5) and 7.5 µM of purified recombinant protein. The reaction mixtures were incubated at 37 °C for 16 h. For HPLC analysis, the reactions were terminated with 100 µL MeOH. Protein was then removed by centrifugation at 17,000 rpm for 15 min. HPLC-measurements were carried out in duplicate on a RP-18 and a Chiralpak Zwix column (+).

Enzyme assays for determination of the kinetic parameters for DMAPP and its analogs MAPP, EAPP and BENZYLPP contained 1 mM L-tryptophan, 5 mM $CaCl_2$ for fungal or $MgCl_2$ for bacterial prenyltransferases, 0.15% (v/v) glycerol, 50 mM Tris-HCl (pH 7.5) and the respective alkyl or benzyl diphosphate in final concentrations of up to 0.5 mM or 1 mM in the case of 5-DMATS with BENZYLPP were incubated at 37 °C in duplicates. For 6-$DMATS_{Sa}$, a protein amount of 5 µg and an incubation time of 30 min were used in the presence of EAPP. For incubation with MAPP or BENZYLPP, the protein amount and incubation time were 10 µg and 60 min. 1 µg 6-$DMATS_{Sv}$ was assayed with DMAPP for 5 min and 10 µg with DMAPP analogs for 60 min. The assays for TyrPT contained 15 µg protein and were incubated for 60 min with EAPP and 90 min with MAPP or BENZYLPP. 10 µg 5-$DMATS_{Sc}$ and an incubation time of 60 min were used for EAPP, 20 µg and 90 min for MAPP. For the reactions with BENZYLPP, 25 µg 5-$DMATS_{Sc}$ and an incubation time of 90 min were used. Kinetic parameters of 5-DMATS were obtained from enzyme assays with 20 µg of purified protein and incubation time of 60 min. The reactions were terminated with 100 µL MeOH and the protein was removed by centrifugation at 17,000 rpm for 15 min. Parameters of Michaelis-Menten kinetics such as K_M and turnover number (k_{cat}) were determined by Lineweaver-Burk, Hanes-Woolf and Eadie-Hofstee plots.

4.5. Enzyme assays for isolation and structure elucidation

Assays for isolation of the enzyme products were carried out in large scales (10 mL) containing 1 mM L-tryptophan, 2 mM MAPP, EAPP or BENZYLPP, 5 mM $CaCl_2$, 0.0%-1.5% (v/v) glycerol, 50 mM Tris-HCl (pH 7.5) and with 2 to 4 mg of purified recombinant protein. After incubation for 16 h at 37 °C, the reaction mixtures were terminated with 10 mL MeOH and precipitated protein was removed by centrifugation at 4,750 rpm for 15 min. The obtained supernatant was then concentrated on a rotating vacuum evaporator to 1 mL for injection in HPLC.

4.6. HPLC analysis and isolation of the enzyme products for structure elucidation

The enzyme products were analyzed on an Agilent series 1200 HPLC (Agilent Technologies Deutschland GmbH, Böblingen, Germany) with a Multospher 120 RP-18 column (250 x 4 mm, 5 µm, C+S-Chromatography Service, Langerwehe, Germany) at a flow rate of 1 mL min^{-1}. Water (solvent A) and methanol (solvent B) were used as solvents for analysis and isolation of the enzyme products. For analysis of the alkylated tryptophan, a linear gradient of 40-100% (v/v) solvent B over 15 min was used. The column was then washed with 100% solvent B for 5 min and equilibrated with 40% solvent B for 5 min. Detection was carried out on a photo diode array detector.

By using the same HPLC equipment and a semipreparative Multospher 120 RP-18 column (250 x 10 mm, 5 µm, C+S-Chromatographie Service, Langerwehe, Germany), the enzyme products were isolated at a flow rate of 2.5 mL min^{-1} and a gradient of 60-100% solvent B in 20-25 min. If necessary, an isocratic step with solvent B before the gradient was included for 5 min. After each run the column was washed with 100% solvent B and equilibrated with 60% solvent B for 5 min.

A much better separation of the L-tryptophan derivatives with different alkylation positions on the indole ring was achieved by using a Chiralpak Zwix column (+) (150 x 3 mm, 3 µm, Chiral technologies Europe, Daicel Group, Illkirch

Cedex, France). This column was used for detailed investigations on the enzyme products in the incubation mixtures (**Figure S1**, Supporting Information) and for separation of the product mixtures, which were not separated by using the semipreparative Multospher 120 RP-18 column mentioned above. Analysis of the enzyme assays and isolation of the products were carried out at a flow rate of 0.5 mL min^{-1} with water (solvent A) and methanol (solvent B) as solvents. An isocratic run with 50% solvent B was used.

4.7. NMR and MS analyses as well as structure elucidation

NMR including two-dimensional HSQC and HMBC spectra were recorded on a JEOL ECA-500 (JEOL Germany GmbH, Munich, Germany) or Bruker Avance-600 spectrometer (Bruker Corporation, Billerica, USA), respectively. The spectra were processed with MestReNova 6.0.2. Chemical shifts were referred to the signals of CD_3OD at δ_H 3.31 and δ_C 49.2 ppm. The isolated compounds were also analyzed by electrospray ionization (ESI-MS) or electron impact mass spectrometry (EI-MS) on a Q-Trap 2000 (Life Technologies Ltd, Paisley PA4 9RF, United Kingdom) and by high resolution electrospray ionization (HR-ESI-MS) or electron impact mass spectrometry (HR-EI-MS) on an Auto SPEC (Waters MS Technology Centre, Manchester, United Kingdom).

Compound Ia: ^{1}H NMR (methanol-d_4, 500 MHz) δ_{ppm} (coupling constant, assignment): 7.60 (dd, *J*=8.1, 0.4 Hz, H-4), 7.15 (d, 0.7, H-7), 7.12 (s, H-2), 6.89 (dd, *J*=8.1, 1.4 Hz, H-5), 5.60 (dtq, *J*=15.1, 6.6, 1.4 Hz, H-2'), 5.51 (dqt, *J*=15.1, 6.3, 1.3 Hz, H-3'), 3.84 (dd, *J*=9.5, 4.0 Hz, H-11), 3.49 (ddd, *J*=15.2, 4.0, 0.6 Hz, H-10), 3.37 (d, *J*=6.7 Hz, H_2-1'), 3.11 (dd, *J*=15.2, 9.5 Hz, H-10), 1.67 (dd, *J*=6.2, 1.4 Hz, H_3-4'); ESI-MS: m/z (intensity) 517.30 [2M+H]$^+$ (100), 539.3 [2M+Na]$^+$ (58), 297.10 [M+K]$^+$ (47), 259.10 [M+H]$^+$ (31), 281.10 [M+Na]$^+$(24); HR-EI-MS: m/z= 258.1322, calcd for $C_{15}H_{18}N_2O_2$ [M]$^+$: 258.1368.

Compound IIa: ^{1}H NMR (methanol-d_4, 500 MHz): 7.60 (d, *J*=8.2 Hz, H-4), 7.16 (d, *J*=0.6 Hz, H-7), 7.13 (s, H-2), 6.90 (dd, *J*=8.2, 1.4 Hz, H-5), 5.63–5.50 (m, H-2'/H-3'), 3.84 (dd, *J*=9.5, 4.0 Hz, H-11), 3.49 (ddd, *J*=15.1, 4.0, 0.6 Hz, H-10), 3.38 (d, *J*=6.0 Hz, H_2-1'), 3.12 (dd, *J*=15.2, 9.5 Hz, H-10), 2.04 (m, H_2-4'), 0.99 (t, *J*=7.5 Hz, H_3-5'); ESI-MS: m/z (intensity) 273.25 [M+H]$^+$ (100), 295.14 [M+Na]$^+$ (87), HR-EI-MS: m/z= 272.1557, calcd. for $C_{16}H_{20}N_2O_2$ [M]$^+$: 272.1525.

Compound IIIa: ^{1}H NMR (methanol-d_4, 500 MHz): 7.60 (dd, *J*=8.2, 0.6 Hz, H-4), 7.25–7.17 (m, H-2'/H-6', H-3'/H-5'), 7.18 (s, H-7, overlaid with H-2'/H-6´, H-3'/H-5'), 7.15–7.11 (m, H-4'), 7.13 (s, H-2' overlaid with H-4'), 6.93 (dd, *J*=8.2, 1.5 Hz, H-5), 4.04 (s, H_2-1'), 3.82 (dd, *J*=9.4, 4.0 Hz, H-11), 3.48 (ddd, *J*=15.2, 4.0, 0.7 Hz, H-10), 3.11 (dd, *J*=15.2, 9.4 Hz, H-10); ESI-MS: m/z (intensity) 295.20 [M+H]$^+$ (100), 589.70 [2M+H]$^+$ (50), 316.92 [M+Na]$^+$ (11); HR-EI-MS: m/z= 294.1368, calcd. for $C_{18}H_{18}N_2O_2$ [M]$^+$: 294.1362.

Compound Ib1: ^{1}H NMR (methanol-d_4, 500 MHz): 7.56 (dd, *J*=7.9, 1.0 Hz, H-4), 7.19 (s, H-2), 7.00 (dd, *J*=7.9, 7.1 Hz, H-5), 6.93 (dd, *J*=7.1, 0.6 Hz, H-6), 5.65 (dtq, *J*=15.1, 6.4, 1.4 Hz, H-2'), 5.57 (dqt, *J*=15.1, 6.2, 1.2 Hz, H-3'), 3.85 (dd, *J*=9.5, 4.0 Hz, H-11), 3.53 (d, *J*=5.2 Hz, H_2-1'), 3.51 (m, H-10, overlaid with H-1'), 3.14 (dd, *J*=15.4, 9.5 Hz, H-10), 1.66 (ddt, *J*=6.0, 1.4, 1.3 Hz, H_3-4'); ESI-MS: m/z (intensity) 259.30 [M+H]$^+$ (100), 281.10 [M+Na]$^+$ (55.8), 539.23 [2M+Na]$^+$ (14), 517.32 [2M+H]$^+$ (8), 297.00 [M+K]$^+$ (6); HR-ESI-MS: m/z= 281.1266, calcd. for $C_{15}H_{18}N_2O_2$ [M+Na]$^+$: 281.1288.

Compound IIb1: ^{1}H NMR (methanol-d_4, 600 MHz): 7.55 (d, *J*=7.8 Hz, H-4), 7.19 (s, H-2), 7.00 (t, *J*=7.5 Hz, H-5), 6.93, (d, *J*=7.0 Hz, H-6), 5.67–5.58 (m, H-2'/H-3'), 3.84 (dd, *J*=9.3, 3.8 Hz, H-11), 3.54 (d, *J*=3.3 Hz, H_2-1'), 3.50 (dd, *J*=15.1, 3.8 Hz, H-10), 3.13 (dd, *J*=15.1, 9.3 Hz, H-10), 2.03 (m, H_2-4'), 0.97 (t, *J*=7.5 Hz, H_3-5'); ^{13}C NMR (methanol-d_4, 150 MHz, deduced from HSQC/HMBC): δ_{ppm}=136.8, 134.1, 128.4, 127.9, 125.1, 124.7, 122.2, 120.2, 117.1, 109.9, 56.5, 35.0, 28.4, 26.1, 13.8; ESI-MS: m/z (intensity) 295.10 [M+Na]$^+$ (100), 273.14 [M+H]$^+$ (29), 545.35 [2M+H]$^+$ (6), 567.39 [2M+Na]$^+$ (6), 311.10 [M+K]$^+$ (3); HR-ESI-MS: m/z=295.1395, calcd. for $C_{16}H_{20}N_2O_2$ [M+Na]$^+$: 295.1422.

Compound IIIb: ^{1}H NMR (methanol-d_4, 500 MHz): 7.58 (dd, *J*=8.1, 0.8 Hz, H-4), 7.26–7.20 (m, H-2'/H-6', H-3'/ H-5'), 7.18 (s, H-2), 7.14 (m, H-4'), 7.01 (t, *J*=7.6 Hz, H-5), 6.92 (dd, *J*=7.2, 0.4 Hz, H-6), 4.20 (s, H_2-1'), 3.83 (dd, *J*=9.4, 4.0 Hz, H-11), 3.50 (dd, *J*=15.1. 4.0 Hz, H-10), 3.13 (dd, *J*=15.1, 9.4 Hz, H-10); ESI-MS: m/z (intensity) 316.90 $[M+Na]^+$ (100), 295.2 $[M+H]^+$ (11.4). HR-EI-MS: m/z=294.1368, calcd. for $C_{18}H_{18}N_2O_2$ $[M]^+$: 294.1339 (as a mixture with **IIIa**).

Compound Ib2: ^{1}H NMR (methanol-*d*$_4$, 500 MHz): 7.56 (dd, *J*=7.8, 1.1 Hz, H-4), 7.20 (s, H-2), 7.03 (t, *J*=7.6 Hz, H-5), 6.98 (d, *J*=7.1 Hz, H-6), 6.12 (ddd, *J*=17.2, 10.3, 6.3 Hz, H-2'), 5.12 (dt, *J*=17.2, 1.6 Hz, H-1'), 5.04 (dt, *J*=10.3, 1.6 Hz, H-1'), 3.89 (m, H-3'), 3.85 (dd, *J*=9.3, 4.1 Hz, H-11), 3.51 (ddd, *J*=15.1, 4.1, 0.9 Hz, H-10), 3.15 (dd, *J*=15.1, 9.3 Hz, H-10), 1.44 (d, *J*=7.0 Hz, H_3-4'); ESI-MS: m/z (intensity) 281.30 $[M+Na]^+$ (100), 259.16 $[M+H]^+$ (63), 539.40 $[2M+Na]^+$ (12), 517.40 $[2M+H]^+$ (8), HR-ESI-MS: m/z=281.1255, calcd. for $C_{15}H_{18}N_2O_2$ $[M+Na]^+$: 281.1266.

Compound IIb2: ^{1}H NMR (methanol-*d*$_4$, 500 MHz): 7.55 (dd, *J*=7.8, 1.1 Hz, H-4), 7.16 (s, H-2), 7.01 (t, *J*=7.5 Hz, H-5), 6.96 (d, *J*=7.5 Hz, H-6), 6.06 (ddd, *J*=17.2, 10.2, 7.6 Hz, H-2'), 5.08 (dt, *J*= 17.2, 1.5 Hz, H-1'), 4.99 (ddd, *J*=10.2, 1.9, 1.0 Hz, H-1'), 3.71 (m, H-11), 3.59 (m, H-3'), 3.41 (m, H-10), 3.04 (dd, *J*=14.9, 8.5 Hz, H-10), 1.86 (m, H_2-4'), 0.89 (t, *J*=7.4 Hz, H_3-5'). Coupling constants of signals observed for H-11, H-3' and H-10 were not determinable, due to low signal intensity; ESI-MS: m/z (intensity) 295.30 $[M+Na]^+$ (100), 273.34 $[M+H]^+$ (18), 568.10 $[2M+Na]^+$ (5), 318.10 $[M+2Na]^+$ (2); HR-ESI-MS: m/z=295.1437, calcd. for $C_{16}H_{20}N_2O_2$ $[M+Na]^+$: 295.1422

Compound Ic: ^{1}H NMR (methanol-*d*$_4$, 500 MHz): 7.50 (s, H-4), 7.27 (dd, *J*=8.4, 0.6 Hz, H-7), 6.96, (dd, *J*=8.5, 1.7 Hz, H-6). Signals at approx. 7.17-7-15 (H-2), 5.67-5.61 (H-2'), 5.54-5.47 (H-3'), 3.86-3.82 (H-11), 3.52-3.48 (H-10), 3,41-3.38 (H-1') 3.12-3.06 (H-10) and 1.69-1.66 (H-4') are overlaid with those of **Ia**. ESI-MS: m/z (intensity) 281.04 $[M+Na]^+$ (100), 259.11 $[M+H]^+$ (10), 517.26 $[2M+H]^+$ (6), HR-EI-MS: m/z=258.1366, calcd. for $C_{15}H_{18}N_2O_2$ $[M]^+$: 258.1368 (in a mixture with **Ia**).

Compound IIc: ^{1}H NMR (methanol-*d*$_4$, 500 MHz): 7.50 (s H-4), 7.27 (d, *J*=8.2 Hz, H-7), 7.15 (s, H-2), 6.97 (dd, *J*= 8.3, 1.6 Hz, H-6), 3,48 (dd, *J*=15.0, 4.6 Hz, H-10), 3.40 (d, *J*=6.6 Hz, H-1'). Signals at approx. 5.66–5.50 (H-2' and H-3'), 3.85-3.82 (H-11), 3.11-3.05 (H-10), 2.07-2.00 (H-4') and 1.00-0.96 (H-5') are overlaid with those of **IIa**; ESI-MS: m/z (intensity) 295.08 $[M+Na]^+$ (100), 273.20 $[M+H]^+$ (7), 567.58 $[2M+Na]^+$ (6), 545.39 $[M+H]^+$(3); HR-ESI-MS: m/z=295.1433, calcd. for $C_{16}H_{20}N_2O_2$ $[M+Na]^+$: 295.1422 (in a mixture with **IIa**).

Compound IIIc: ^{1}H NMR (methanol-*d*$_4$, 500 MHz): 7.60 (dd, *J*=1.5, 0.7 Hz, H-4), 7.27 (dd, *J*=8.3, 0.5 Hz, H-7), 7.16 (s, H-2), 6.97 (dd, *J*=8.3, 1.6 Hz, H-6), 4.05 (s, H_2-1'), 3.84 (m, H-11), 3.51 (m, H-10). Signals at approx. 7.25–7.17 (H-2'/H-6', H-3'/H-5'), 7.14-7.10 (H-4'), 3.12-3.06 (H-10), are overlaid with those of **IIIa**; ESI-MS: m/z (intensity) 295.14 $[M+H]^+$ (100), 589.19 $[2M+H]^+$ (31), 317.37 $[M+Na]^+$ (6); HR-EI-MS: m/z=294.1368, calcd. for $C_{18}H_{18}N_2O_2$ $[M]^+$: 294.1339 (as a mixture with **IIIa**).

4.8. Docking studies

All calculations were carried out using FRED (34) and conformations of tryptophan and the four donor molecules were generated with OMEGA (OMEGA 2.5.1.4: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com. Hawkins, P.C.D.; Skillman, A.G.; Warren, G.L.; Ellingson, B.A.; Stahl, M.T.). For receptor preparation, the homology model of 5-DMATS was processed with the apopdb2receptor-tool (part of the FRED docking suite) in order to determine the docking volume. The five molecules, *i.e.*, tryptophan and the four donor molecules DMAPP, MAPP EAPP and BENZYLPP, were docked independently, storing the best 10,000 poses of each for further processing.

The first percentile of the stored poses was selected for visual inspection. Among these poses, the most reasonable ones by physicochemical criteria were selected as starting points for molecular dynamics (MD) simulations.

4.9. Molecular dynamics studies

The three-dimensional model structure of 5-DMATS used for docking and the selected docking poses of pairs of substrate and donor were processed further using MOE (Molecular Operating Environment (MOE) 2010.10.

Quebec: Chemical Computing Group; 2010): Protonation states were calculated with the Protonate3D routine within MOE and visually inspected for plausibility. AMBER atom types were assigned to protein, substrate and donor atoms.

The MD simulations were run with the Amber14 software suite (University of California, San Francisco). The force field parameters were determined using the antechamber program (within the Amber14 suite). Amber coordinate, parameter and topology files were generated by xleap and an octahedral explicit water box (based on the TIP3P water model (35) was constructed 10 Å away from the protein. The resulting systems were minimized, heated from 100 K to 300 K over 20 ps at constant number of particles, volume and temperate (NVT) and equilibrated at 300 K for 100 ps at constant pressure (NPT) with unrestrained water molecules and restrained substrate and donor molecules. One more minimization step and subsequent heating from 100 K to 300 K for 20 ps (NVT) followed by five separate equilibration steps (four steps of 100 ps each and a final step of 2 ns at 300 K (NPT)) were performed while lowering the restraints applied to protein, substrate and donor with each step (unrestrained system at the final equilibration step). The productive simulation was carried out for 5 ns at 300 K (NVT) and 2 fs time step, storing the coordinates every picosecond. All simulations were carried out with the pmemd.cuda module of the Amber14 suite on four GPUs.

The simulations were visualized with VMD. (36) Dynamic trajectories analysis and geometric data extraction was performed with cpptraj (Amber14 suite). Graphical representations of the simulated complexes were prepared using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.).

5. Acknowledgments

This work was supported by a grant from the Deutsche Forschungsgemeinschaft (Grant Li844/4-1 to S.-M. Li).

Julia Winkelblech and Jakub Gunera are financially supported by the LOEWE program of the State of Hessen (SynMikro to S.-M. Li and P. Kolb).

We thank Dr. Laufenberg and Lena Ludwig (Philipps-Universität Marburg) for acquiring mass spectra and for the synthesis of DMAPP, respectively.

References

1. J. Clardy, C. Walsh, *Nature* **2004,** *432,* 829-837.

2. D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2012,** *75,* 311-335.

3. C. T. Walsh, M. A. Fischbach, *J. Am. Chem. Soc.* **2010,** *132,* 2469-2493.

4. P. Wiemann, N. P. Keller, *J. Ind. Microbiol. Biotechnol.* **2014,** *41,* 301-313.

5. S. D. Bentley, K. F. Chater, A. M. Cerdeno-Tarraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C. H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M. A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B. G. Barrell, J. Parkhill, D. A. Hopwood, *Nature* **2002,** *417,* 141-147.

6. H. Ikeda, J. Ishikawa, A. Hanamoto, M. Shinose, H. Kikuchi, T. Shiba, Y. Sakaki, M. Hattori, S. Omura, *Nat. Biotechnol.* **2003,** *21,* 526-531.

7. Y. Ohnishi, J. Ishikawa, H. Hara, H. Suzuki, M. Ikenoya, H. Ikeda, A. Yamashita, M. Hattori, S. Horinouchi, *J. Bacteriol.* **2008,** *190,* 4050-4060.

8. A. A. Brakhage, *Nat. Rev. Microbiol.* **2013,** *11,* 21-32.

9. L. Heide, *Curr. Opin. Chem. Biol.* **2009,** *13,* 171-179.

10. S.-M. Li, *Nat. Prod. Rep.* **2010,** *27,* 57-78.

11. X. Yu, S.-M. Li, *Methods Enzymol.* **2012,** *516,* 259-278.

12. S. Takahashi, H. Takagi, A. Toyoda, M. Uramoto, T. Nogawa, M. Ueki, Y. Sakaki, H. Osada, *J. Bacteriol.* **2010,** *192,* 2839-2851.

13. X. Yu, Y. Liu, X. Xie, X.-D. Zheng, S.-M. Li, *J. Biol. Chem.* **2012,** *287,* 1371-1380.

14. N. Steffan, I. A. Unsöld, S.-M. Li, *Chembiochem* **2007,** *8,* 1298-1307.

15. A. Kremer, S.-M. Li, *Appl. Microbiol. Biotechnol.* **2008,** *79,* 951-961.

16. S. Subramanian, X. Shen, Q. Yuan, Y. Yan, *Process Biochem.* **2012,** *47,* 1419-1422.

17. T. Ozaki, M. Nishiyama, T. Kuzuyama, *J. Biol. Chem.* **2013,** *288,* 9946-9956.

18. J. Winkelblech, S.-M. Li, *Chembiochem.* **2014,** *15,* 1030-1039.

19. M. Liebhold, X. Xie, S.-M. Li, *Org. Lett.* **2012,** *14,* 4884-4885.

20. M. Liebhold, S.-M. Li, *Org. Lett.* **2013,** *15,* 5834-5837.

21. A. Fan, H. Chen, R. Wu, H. Xu, S.-M. Li, *Appl. Microbiol. Biotechnol.* **2014,** *98,* 10119-10129.

22. M. Liebhold, X. Xie, S.-M. Li, *Org. Lett.* **2013,** *15,* 3062-3065.

23. A. Kremer, L. Westrich, S.-M. Li, *Microbiology* **2007,** *153,* 3409-3416.

24. A. Kremer, S.-M. Li, *Microbiology* **2010,** *156,* 278-286.

25. U. Metzger, C. Schall, G. Zocher, I. Unsöld, E. Stec, S.-M. Li, L. Heide, T. Stehle, *Proc. Natl. Acad. Sci. U. S. A* **2009,** *106,* 14309-14314.

26. M. Jost, G. Zocher, S. Tarcz, M. Matuschek, X. Xie, S.-M. Li, T. Stehle, *J. Am. Chem. Soc.* **2010,** *132,* 17849-17858.

27. J. M. Schuller, G. Zocher, M. Liebhold, X. Xie, M. Stahl, S.-M. Li, T. Stehle, *J. Mol. Biol.* **2012,** *422,* 87-99.

28. X. Yu, G. Zocher, X. Xie, M. Liebhold, S. Schütz, T. Stehle, S.-M. Li, *Chem. Biol.* **2013,** *20,* 1492-1501.

29. A. B. Woodside, Z. Huang, C. D. Poulter, *Org. Synth.* **1988,** *66,* 211-215.

30. T. Ozaki, M. Nishiyama, T. Kuzuyama, *J. Biol. Chem.* **2013,** *288,* 9946-9956.

31. A. Šali, T. L. Blundell, *J. Mol. Biol.* **1993,** *234,* 779-815.

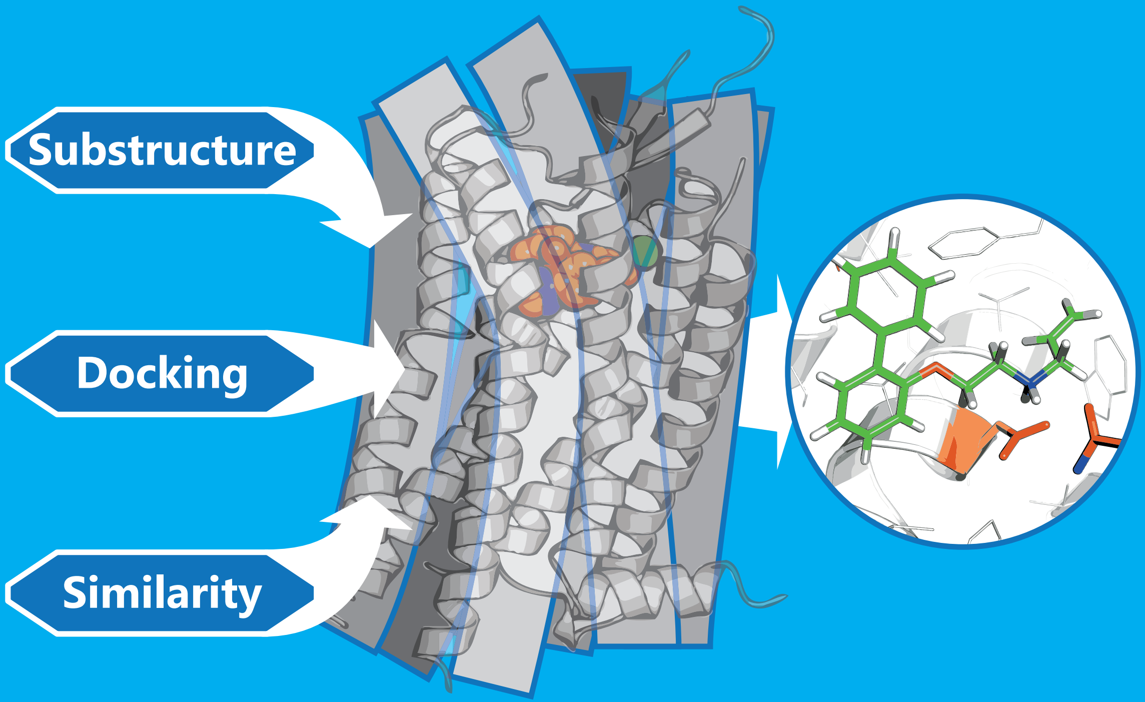
32. M. Y. Shen, A. Sali, *Protein Sci.* **2006,** *15,* 2507-2524.

33. B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, M. Karplus, *J. Comput. Chem.* **2009,** *30,* 1545-1614.

34. M. McGann, *J. Chem. Inf. Model.* **2011,** *51,* 578-596.

35. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983,** *79,* 926-935.

36. W. Humphrey, A. Dalke, K. Schulten, *J. Mol. Graph.* **1996,** *14,* 33-38.



ACS Medicinal Chemistry Letters **2017**, *8*, 451-485.

Part 4

Part 4 is composed of a joint study between the lab of Peter Kolb and Jillian Baker, where we combined the popular three- and two-dimensional ligand discovery approaches – substructure and similarity search and docking – in order to expand the chemical matter around previously described novel scaffolds for the β_2-adrenergic receptor. In this study, we describe the discovery of analoguous compounds via a large scale high-throughput virtual screening cascade and their thorough experimental characterization in alignment with the SAR derived from their putative binding geometries.

The author list is the following (by contribution order): Schmidt D., Gunera J, Baker J. G. and Kolb P. Peter did the original similarity and substructure searches and docking calculations. Denis and I acquired compounds, prepared the assay-ready formats and supervised initial affinity measurements. Jillian performed the pharmacological experiments and data analysis. Peter, Denis and I discussed the data in accordance to the generated putative binding geometries in order to establish an SAR profile.

Similarity- and substructure-based development of β_2-adrenergic receptor ligands based on unusual scaffolds

Denis Schmidt[†,‡], Jakub Gunera[†], Jillian G. Baker[¶] and Peter Kolb[†,*]

[†]Department of Pharmaceutical Chemistry, Philipps-University, Marburg, Hesse, 35032, Germany

[‡]Institute of Pharmaceutical and Medicinal Chemistry, Heinrich-Heine-University, Düsseldorf, 40225, Germany

[¶]Cell Signalling, School of Life Science, Queen's Medical Centre, University of Nottingham, Nottingham NG7 2UH, U.K.

Keywords:

β_2-adrenergic receptor, similarity searches, docking, SAR-by-catalog

Abstract

The β_2-adrenergic receptor (β_2AR) is a G protein-coupled receptor (GPCR) and a well-explored target. Here, we report the discovery of 13 ligands – ten of which are novel – of this particular GPCR. They have been identified by similarity- and substructure-based searches using multiple ligands – which were described in an earlier study – as starting points. Of note, two of the molecules used as queries here distinguish themselves from other β_2AR antagonists by their unique scaffold. The molecules described in this work allow us to explore the ligand space around the previously reported molecules in greater detail, leading to insights into their structure-activity relationship. We also report experimental binding and selectivity data and putative binding modes for the novel molecules.

1. Introduction

The membrane receptors of the G protein-coupled receptor (GPCR) family are flexible heptahelical bundles transferring signals from the outside to the inside of a cell. This is achieved by a conformational change of the receptor upon binding of a signaling molecule to a cavity located at the extracellular end between the seven helices. GPCRs are expressed in almost all tissues (1), and it is thus not surprising that approximately 1/3 of present-day drugs interact with a GPCR. (2) Among these receptors, the β_2-adrenergic receptor (β_2AR) is considered a prototypical representative, and has been investigated for more than 60 years. It was also the first pharmacologically relevant GPCR to succumb to crystallization in 2007. (3,4)

In a previous work (5), we have identified six ligands (originally labeled **1-6**, and referred to as **Q1-Q6** in this work to avoid confusion, **Chart S1**) of the β_2AR through *in silico* docking studies, with affinities ranging from 9 nM to 3.2 μM. Notably, these included two molecules (**5** and **6** in (5), denoted as **Q5** and **Q6**, respectively, in the following) that did not follow the classical adrenaline-based scaffold. (6) This was remarkable, as nobody had discovered these scaffolds earlier, despite more than six decades of medicinal chemistry in this area. Building upon the discovery of the six ligands, we wanted to expand chemical space around them. In particular, we wanted to investigate the two ligands with unusual scaffolds by employing *in silico* similarity and substructure searches in the ZINC (7) database. Candidate molecules identified in either way were then docked into the β_2AR, in order to ascertain that their binding modes were consistent. Here we report the results

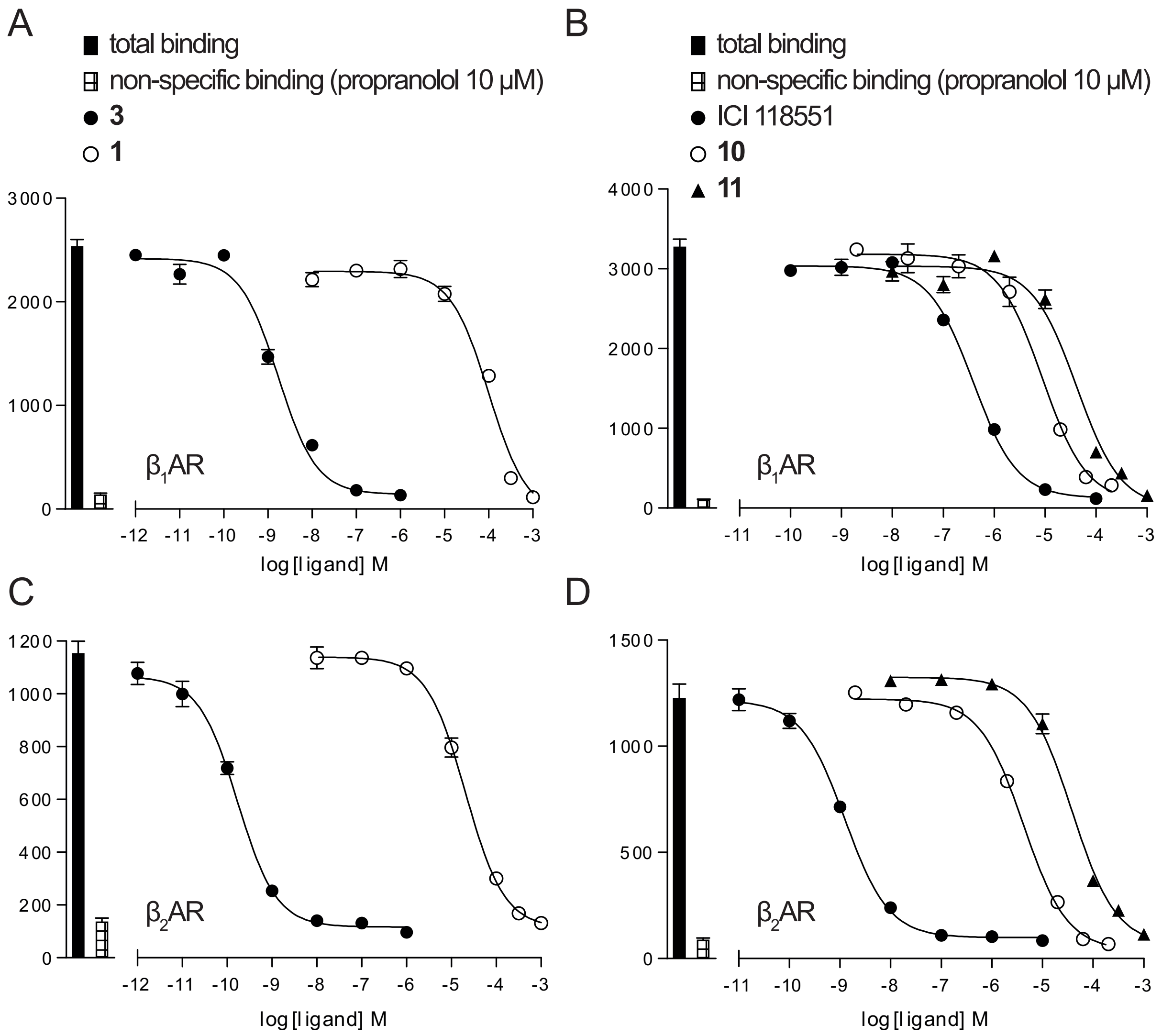


Figure 1: Inhibition of [^{3}H](-)CGP 12177 whole cell binding to **A**, **B** CHO-β_1 cells and **C**, **D** CHO-β_2 cells in response to **A** , **C** **3** and **1** and **B**, **D** ICI 118551, **10** and **11**. Bars represent total and non-specific binding and data points are mean ± SEM of triplicate determinations. The concentration of [^{3}H](-)CGP 12177 used in these experiments was **A**, **C** 0.58 nM and **B**, **D** 0.44 nM and they are representative of **A** 4, **B** 5, **C** 5 and **D** 5 separate experiments.

of this combined ligand- and structure-based screen, which also provides insights into the structure-activity relationship (SAR) of molecules **Q5** and **Q6** and their derivatives.

2. Results

The similarity screen amongst the 8.5 million molecules of the ZINC database resulted in 6,363 molecules, which were distributed across the six query molecules as shown in **Table S1**. From the substructure-based screen, approximately 653,000 hits emerged. Duplicates were removed from both sets. After docking, 5,838 and 587,099 molecules remained, respectively, and the top-scoring 500 of each run were visually inspected. After weeding out molecules with artificially inflated scores due to the absence of corrective terms in present-day scoring functions, *e.g.*, unfavorable desolvation contributions or unsatisfied hydrogen-bond donors, during this inspection, we were left with eight and nine molecules from the similarity and substructure searches, respectively. These were acquired from their respective vendors for further experimental testing (**Table S5**). Three compounds (**1**, **2**, and **3**) contained a biaryl moiety and a charged amine and thus resembled the classical motif of a β_2-binder. Indeed, a thorough literature search revealed that these compounds had been described before (**Table 1**; by the time of selection, these compounds had not been

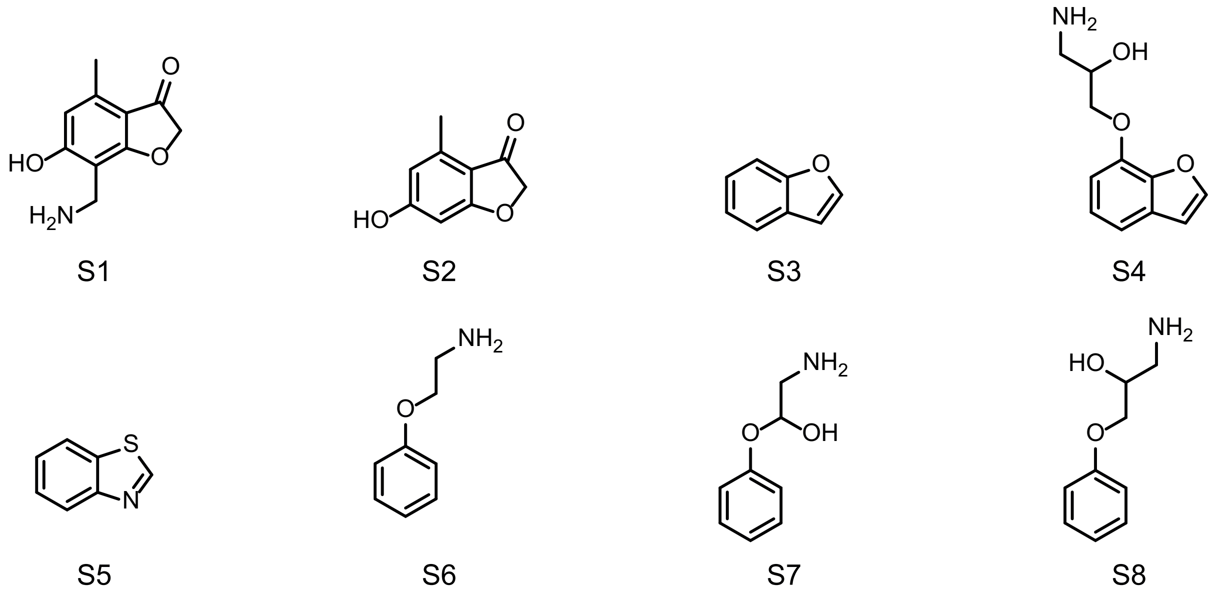


Chart S2: The eight substructures, based on the ligands of ref. (5), used for screening in this study.

annotated in ChEMBL (8)). To analyze the selectivity of the compounds, we also evaluated them against the closely related β_1AR. The efficacy of all compounds was further evaluated in a functional assay.

Several of the compounds identified in this work inhibited [^{3}H](-)CGP 12177 whole cell binding (**Table 1**; see Supporting Information for assay validation and **Table S3** for inactive compounds). This assay also demonstrated that compound **3** had very high affinity (pK_D 9.01 at β_1AR and pK_D 10.45 at β_2AR) and was therefore 28-fold β_2-selective (**Figure 1A** and **C**, **Table 1**). While the remaining compounds had relatively poor affinity in comparison to **3**, many of them, *e.g.*, **1**, **2**, **10**, **11** and **13**, inhibited [^{3}H](-)CGP 12177 binding to yield measurable affinity values (**Figure 1B** and **D**, **Table 1**).

Next, characteristics of ligands were examined in a functional assay, namely CRE-gene transcription. The ability of ligands to stimulate a response (intrinsic efficacy) was assessed but also, given that the affinity of many of the ligands to inhibit [^{3}H](-)CGP 12177 binding were at the very limit of the *binding* assay, the ability of ligands to inhibit *functional* responses was also evaluated, thus giving a totally independent measure of affinity from that achieved in the binding assay.

Except for compound **3**, no other compound stimulated a measurable response (n=4-5 for each compound) in this assay (see Supporting Information for more details and assay validation). However, several compounds antagonized the cimaterol response to give a parallel shift of the cimaterol concentration response curve and thus yield measurable K_D values (**Figure S1**, **Table S2**). For some compounds, *e.g.*, **1**, **2**, and **13**, this gave selectivity values similar to those obtained in the binding assay. For other compounds, *e.g.*, **16** and **17**, no rightward shift of the cimaterol response was observed, suggesting no inhibition at the maximum concentration possible (100 µM in each case). For few of the ligands, the highest concentration possible caused a marked fall in CRE-SPAP production to below basal in a manner more consistent with toxicity, cell death or assay interference, rather than receptor-mediated inverse agonism (see Supporting Information for full details). In these instances, compound concentrations used to inhibit cimaterol responses were reduced until such a time as the reduction in basal was minimal. An example of this was compound **10**, which reduced basal at the maximum concentration of 20 µM but not at 2 µM (see Supporting Information). At 2 µM, **10** was still able to cause a rightward shift of the cimaterol concentration response curve at the β_2AR, but not the β_1AR, consistent with its β_2-selectivity. The fall from maximum of the concentration response to cimaterol (most likely because the assay is at the limit of its capability) means that an apparent K_D is reported (calculated

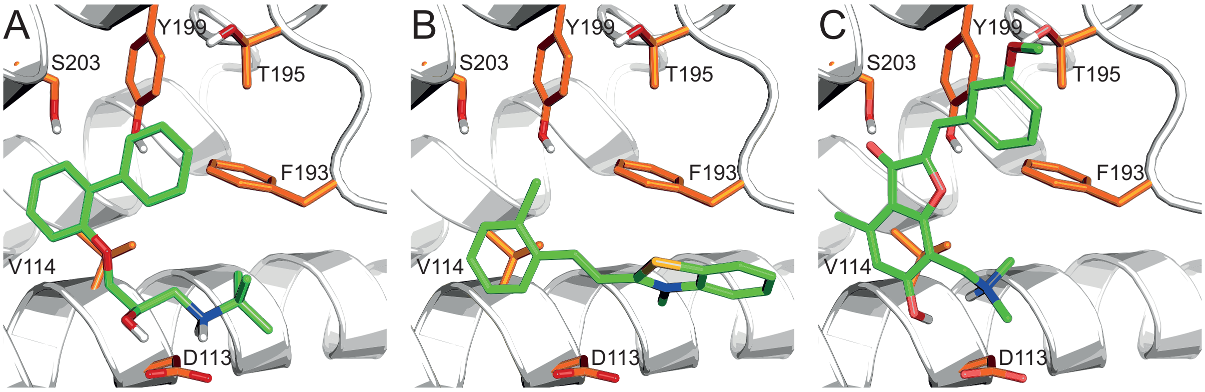


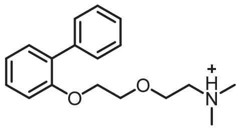
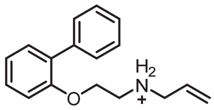
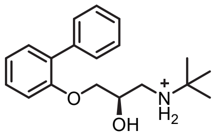
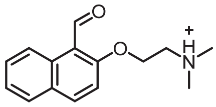
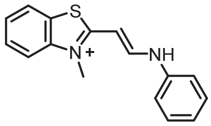
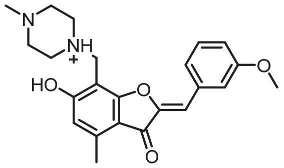
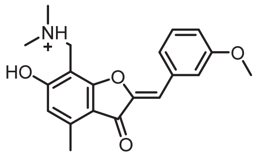
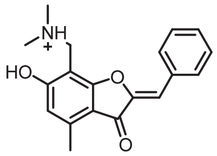
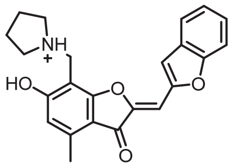
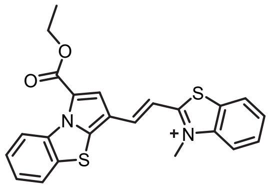
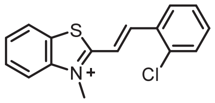
Figure S5: Docking poses for selected compounds. The β2AR is shown in gray stick representation. Residues discussed in the text are labeled and shown with colored heteroatoms. Selected residues in TM6 and TM7 (including Phe289$^{6.51}$ and Phe290$^{6.52}$) are hidden for clarity. Ligands are shown in orange stick representation. Perspective as in ref. (5) for comparability. **A 3**, **B 11**, **C 7**.

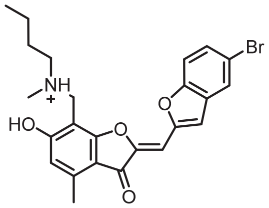
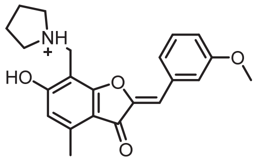
from the shift of the lower part of the curve where the lines are parallel), this apparent K_D is however similar to the K_D values obtained from the binding assay, confirming that this is receptor-mediated and β_2-selective.

Compound **3** on its own stimulated a partial agonist response at both the β_1- and β_2AR. This response was inhibited by CGP 20712A in the CHO-β_1-cells with high affinity and by ICI 118551 in the CHO-β_2-cells (**Figure S2**, **Table S4**). Furthermore, **3** was able to inhibit the cimaterol responses in both cell lines in a manner consistent with that of a partial agonist (**Figure S2**, **Table S2**). Finally, **3** inhibited the response to fixed concentrations of cimaterol in both cell lines in a manner consistent with competition at a single receptor conformation (9) (**Figure S1** and Supplementary Procedures for full details).

Altogether, the high affinity of CGP 20712A and ICI 118551 for the CHO-β_1 and CHO-β_2 cells confirm the presence of the β_1- and β_2AR in the respective cell lines. Several of the compounds (*e.g.*, **16** and **17**) did not interact with the receptors in either the binding assay or functional assay up to the maximum concentration possible for the compounds (20-100 μM). Of the molecules with novel scaffolds, **10** and **11** show the highest affinities at pK_D values of 6.05 and 5.31, respectively, for the β_2AR and are thus in a range comparable to those of the established compounds **1** and **2**. These compounds did not induce a functional response in the receptor and are therefore neutral antagonists. However, we emphasize that the outcome of a virtual screening campaign in the manner conducted here is the prediction of binding, not efficacy. Of the novel compounds, **13** exhibited affinity in the binding as well as in the functional assay with low micromolar activity.

Table 1: Affinity (pK$_D$ values) and β_2-selectivity for compounds as measured by [^{3}H](-)CGP 12177 whole cell binding to CHO-β_1 and CHO-β_2 cells. Values are mean ± SEM of n separate experiments.

ID	Structure	β_2AR pK$_D$		n	β_1AR pK$_D$		n	β_1/β_2[a]
1[c]		5.42	±0.14	5	4.34	±0.07	4	12.0
2[c]		5.58	±0.06	6	4.56	±0.06	6	10.5
3[d]		10.45	±0.05	8	9.01	±0.04	5	27.5
4		4.63	±0.07	5	4.01[b]	±0.05	5	4.2
5		4.41	±0.08	3	3.59[b]	±0.1	3	6.6
6		4.76	±0.09	5	4.58	±0.03	5	1.5
7		4.66	±0.16	5	4.35	±0.04	4	2
8		4.60 [b]	±0.11	4	4.33 [b]	±0.05	4	1.9
9		4.84 [b]	±0.13	4	4.42 [b]	±0.11	4	2.6
10		6.05	±0.11	6	5.51	±0.07	6	3.5
11		5.31	±0.12	6	4.86	±0.05	5	2.8

12		4.75[b]	±0.12	5	n.c.		4	
13		5.26	±0.06	6	4.45	±0.04	5	6.5
	ICI 118551	9.61	±0.05	5	6.74	±0.01	5	741
	CGP 20712A	5.84	±0.10	5	8.96	±0.13	4	0.0008

[a]Selectivity: β_1/β_2=$K_D(\beta_1)/K_D(\beta_2)$

[b]apparent K_D values: here the maximum concentration of the compound was not sufficient to fully inhibit specific binding; however, the majority of specific binding was inhibited allowing an apparent measure of affinity.

[n.c.]For ligands with less than 50% inhibition of specific binding, the IC_{50} value could not be determined and thus a K_D value could not be calculated (n.c.)

[c]US 20090163545

[d]Antiarrythmic pharmaceutical (Bipranol/Berlafenone), Arzneimittel-Forschung **1992**, *42*, 289-291.

The more traditional biaryl compounds **1**, **2**, and **3** display the highest affinities at the β_2AR, as was to be expected. In particular, compound **3** was confirmed as a very high affinity partial agonist at both receptors, but with some β_2AR selectivity. At the β_2AR, the affinity measured by binding (pK_D 10.45) and the affinity measured as antagonism of the cimaterol response (pK_D 10.74) are very similar, confirming the very high affinity ligand-receptor interaction. The partial agonist was itself antagonized by ICI 118551 (yielding a similar pK_D for ICI 118551 as that for antagonism of the cimaterol response), confirming that signaling is indeed occurring via the β_2AR. Compound **3** is therefore a very high affinity, weak partial agonist of the human β_2AR. Moreover, **3** was found to be a partial agonist of the β_1AR, with the agonist response occurring through the primary catecholamine conformation of the receptor (see Supplementary Results).

These three molecules, **1**, **2** and **3**, were selected by similarity to compounds **Q2**, **Q3,** and **Q4**, all of which contain a biaryl moiety. Not unexpectedly, these hits not only show high affinities but also highest similarities to known (again exclusively biaryl-containing) compounds that are annotated in the ChEMBL database (**Table S6**). This is encouraging with respect to the performance of similarity screening methods and the value of docking in identifying such compounds. However, it also strongly emphasizes the need for methods that allow for scaffold-hopping to fully explore the ligand space of a target.

By reducing the biaryl scaffold to a 2-ethoxy-ethylamine (**S6** in **Chart S2**) for the substructure search, two more substances, **4** and **14**, were identified. Compound **4** showed two-digit micromolar affinity, whereas the inhibition by **14** was so weak that no reliable affinity value could be calculated. Interestingly, in **14** the nitrogen matched in the substructure search is the one in the benzoxazine portion, not the exocyclic amine.

Turning to the hits derived from reference molecules **Q5** and **Q6**, we note that they show a much lower Tanimoto similarity of approximately 0.3 and below (when compared to molecules from the ChEMBL database using ECFP4 fingerprints) than the other hits reported in ref. (5) (**Table S6**). This is in line with the fact that these compounds are not based on the classical propanolamine scaffold and underlines the structural novelty of these two scaffolds.

Starting from the benzothiazole-based compound **Q5**, four molecules were identified with benzothiazole (**5**, **10**, **11, 15**) and two with benzimidazole (**16**, **17**) motifs. Of these, all benzothiazole-containing molecules except **15** show affinity towards the β_2AR in the micromolar range. Docking poses indicate that the orientation of the benzothiazole ring is comparable to the one of **Q5**, with a polarized methyl group interacting with Asp113$^{3.32}$ (**Figure S5**, **Figure S6**). The benzimidazole and benzoxazole compounds **16** and **17** show no activity in our assay. These compounds might be more sterically hindered in the vicinity of the positively charged nitrogen atom, in particular compound **16**. Furthermore, the different polarity of the ring system, owing to the variation of the heteroatoms, might render the predicted interaction with Asp113$^{3.32}$ less likely.

Six additional compounds could be identified on the basis of the parent molecule **Q6**. All these molecules (**6**, **7**, **8**, **9**, **12**, and **13**) share a benzofuran-based moiety, independent of whether they originated from the substructure or the similarity search. This moiety, namely, a 3-oxo-4-methyl-6-hydroxy-benzofuran, is present in the parent molecule **Q6**, too, and can thus be considered a "stable scaffold" in terms of SAR. All molecules display affinity, with pK_D values varying between 5.26 and 4.6. Interestingly, **8**, which is the substance with the weakest affinity in this set, differs from **7** only by a methoxy group, which is absent in **8**. This methoxy group could act as an acceptor, which is also present in all remaining molecules of this series as (benzo-)furan or methoxy group. The role of this group is not clearly evident from the docking predictions, but an interaction with Thr195^{ECL2} seems to be the most likely explanation (**Figure S5**, **Figure S6**). Furthermore, the docking poses indicate a binding mode of this scaffold, which resembles the key interactions seen in biaryl-based compounds. The benzofuran scaffold forms interactions with Phe193$^{45.52}$, Phe289$^{6.51}$, Phe290$^{6.52}$, and Val114$^{3.33}$. The hydroxy group at position 6 forms an additional hydrogen bond to Asp113$^{3.32}$, while the ketone serves as acceptor for a hydrogen bond from Ser203$^{5.42}$. A second aromatic moiety is attached at position 2, interacting with Tyr199$^{5.38}$, Tyr308$^{7.35}$, and, presumably, Thr195^{ECL2}. An increased size of the aromatic system appears to be detrimental for affinity (methoxyphenyl in **13** vs benzofuran in **9**). The charged amine in the pyrrolidine moiety is expected to form a salt bridge with Asp113$^{3.32}$.

3. Materials and Methods

Substructure queries (**Chart S2**) were manually derived from the original hits. Substructure and similarity searches were run on the ZINC (7) database and docked to the β_2AR (PDB 2RH1), as previously described. (5) [^{3}H](-)CGP 12177 whole cell binding and CRE-SPAP production assays were run using CHO-K1 cells expressing either the human β_1AR or the human β_2AR as previously described. (10,11) See Supporting Information for detailed descriptions of experimental procedures.

4. Discussion and conclusions

We have elaborated on six previously identified novel binders of the β_2AR through SAR-by-catalog. Using similarity and substructure searches followed by a docking assessment of the interactions of each compound and the receptor, 13 ligands of the β_2AR were verified experimentally. Ten of these molecules are indeed novel ligands for the receptor, while the remaining three turned out to have been described before. Based on this data, several conclusions can be drawn.

First, the benzofuran scaffold of compound **Q5** and the benzothiazole scaffold of compound **Q6** in ref. (5) indeed constitute novel chemotypes with derivatization potential for this receptor. Especially the benzofuran series showed a consistent SAR that is in agreement with the predicted binding modes. This study can thus also provide retrospective evidence that the predicted binding modes are indeed very likely correct. The affinities of the novel compounds are not comparable with those of highly optimized adrenaline- or biaryl-based scaffolds. The latter are exemplified by **Q1** with an affinity of 9 nM and **3** with its pK_D of 10.74. However, the novel compounds can serve as unprecedented starting points for further optimization.

Second, that the combination of similarity- and substructure-based searches with protein-structure-based docking constitutes a powerful combination. This is manifest in the quite high hit rate (more than 75 % of the molecules bind with an affinity below 100 µM) and the fact that we (re)discovered a molecule with an affinity of only 35 pM. This compound is also known as *bipranol* or *berlafenone* – an anti-arrythmia drug.

In terms of selectivity, most of the compounds displaying an affinity are mildly selective towards the β_2AR. Again, **3** takes the lead here at 28-fold selectivity for the β_2AR. While other compounds such as **1** and **2** still have at least ten-fold preference towards the β_2AR, all values are far below 100-fold, which is considered a ratio that is significant enough to call a compound "selective". Moreover, highly optimized compounds such ICI 118551 show affinity ratios that are closer to 1000-fold. Interestingly, the top three compounds in terms of selectivity all belong to the biaryl cluster of molecules.

Not unexpectedly, most of the compounds with measurable affinity (with the exception of **3**), turned out to be neutral antagonists in the functional assay. This is consistent with what we have seen in our previous study (5) and the fact that we have been docking to an inactive conformation of the receptor. (3,4)

Future studies will show to which affinities the novel scaffolds can be optimized. It is also encouraging to have confirmed that unbiased computational methods can present us with novel molecules, even for target proteins as well-investigated as the β_2AR.

5. Acknowledgments

We thank the Shoichet lab at the University of California, San Francisco for input when choosing some of the compounds.

Daniel M. Rosenbaum, then at the Kobilka lab at Stanford, is acknowledged for initial assays to categorize some of the compounds.

We thank Richard Proudman for his technical assistance with tissue culture.

P.K. wishes to acknowledge the German Research Foundation DFG for Emmy Noether fellowship KO4095/1-1. J.G.B., P.K., and all lab members are furthermore participating in COST Action CM1207 GLISTEN.

References

1. Regard, J. B.; Sato, I. T.; Coughlin, S. R. Anatomical profiling of G protein-coupled receptor expression. *Cell* **2008,** *135,* 561-571.

2. Salon, J. A.; Lodowski, D. T.; Palczewski, K. The significance of G protein-coupled receptor crystallography for drug discovery. *Pharmacol. Rev.* **2011,** *63,* 901-937.

3. Rosenbaum, D. M.; Cherezov, V.; Hanson, M. A.; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S.; Choi, H.-J.; Yao, X.-J.; Weis, W. I.; Stevens, R. C.; Kobilka, B. K. GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science* **2007,** *318* (5854), 1266-1273.

4. Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S.; Choi, H.-J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High Resolution Crystal Structure of an Engineered Human β2-Adrenergic G protein-Coupled Receptor. *Science* **2007,** *318* (5854), 1258-1265.

5. Kolb, P.; Rosenbaum, D. M.; Irwin, J. J.; Fung, J. J.; Kobilka, B. K.; Shoichet, B. K. Structure-based discovery of beta2-adrenergic receptor ligands. *Proc. Natl. Acad. Sci. U. S. A.* **2009,** *106* (16), 6843-6848.

6. Overington, J.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug. Discov.* **2006,** *5* (12), 993-996.

7. Irwin, J. J.; Shoichet, B. K. ZINC - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005,** *45* (1), 177-182.

8. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, N.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Res* **2014,** *42* (D1), D1083-D1090.

9. Baker, J. G. Sites of action of β-ligands at the human β1-adrenoceptor. *J. Pharmacol. Exp. Ther.* **2005,** *313,* 1163-1171.

10. Baker, J. G. The selectivity of β-adrenoceptor antagonists at the β1, β2 and β3 adrenoceptors. *Br. J. Pharmacol.* **2005,** *144,* 317-322.

11. Baker, J. G.; Proudman, R. G. W.; Hill, S. J. Identification of key residues in transmembrane 4 responsible for the secondary, low affinity conformation of the human β1-adrenoceptor. *Mol. Pharmacol.* **2014,** *85,* 811-829.

Conclusions

Novel tools for computer-aided drug discovery and design

Corpora non agunt nisi fixata.

Paul Ehrlich, 1854-1915

Paul Ehrlich, a German physician, scientist and Nobel-Prize laureate (1908), reasoned that a substance is not biologically active unless it is bound to a biological macromolecule. In the present days, this maxim seems trivially obvious: The drug discovery and design pipeline, the evolution of a drug molecule, starts with the identification of a target – the biological macromolecule. A compound is then searched for that binds to this macromolecule both tightly and selectively and, equally importantly, efficaciously, *i.e.* showing the desired effect. In the days of Paul Ehrlich, without the knowledge gained from over 100,000 crystal structures, without isothermal calorimetry and mass spectroscopy, this maxim exudes a great intuition and pioneering understanding of Nature. It is therefore not surprising that affinity, activity, potency, free energy of binding, or simply the tightness of binding of a small molecule ligand to its target macromolecule is one of the key figures of drug development efforts.

The tools and concepts presented in this thesis are not capable of calculating or predicting this figure explicitly. They are not meant to do so. In case of molecular similarity and **RedFrag** the tightness of binding is abstracted to a probability or score value which is assigned to a given molecule based on the similarity towards a potent, tightly binding, query compound. The exploitation of the similarity dogma – similar molecules share similar bioactivity – replaces the rigorous calculation of binding affinity.

Even more abstractly, but still along the same reasoning of ligand-receptor binding, **PrenDB**, was developed to predict if and at which position a molecule can be prenylated, in other words diversified. Prenylated molecules are meant to be entry points of chemical space only sparsely occupied by reaction products of conventional chemical synthesis, eventually resulting in more diverse and more potent compounds. The prenyl moiety itself can lead to an increase of the binding affinity of the molecule as it increases the lipophilicity of the compound. This molecular obesity, although discouraged in the light of the more favorable enthalpic optimization, opens new routes to hit discovery and optimization.

1. Do we need yet another molecular similarity method?

Well, absolutely yes! Truly, there are many concepts and methods that exploit the molecular similarity in one way or another, always under the assumption that the similarity dogma – similarity property principle – is valid. But, since the molecular similarity, and similarity in general, lies in the eye of the beholder, there are many possibilities how "similarity" can be established and put into numbers. The variety of available fingerprints, *e.g.*, ECFP, FCFP, MACCS keys, OpenEye's Lingo fingerprint, RDKit's topological torsions and atom-pair fingerprint, shows that molecular similarity is an inspiring and eclectic concept. However, the prospect of finding a compound of similar or better affinity is, although certainly useful, not entirely satisfying. It seems much more attractive to find a compound that is *different* from the query *and* of similar or better affinity. **RedFrag** was designed with this idea of scaffold hopping in mind.

Our idea and central pivot point within the RedFrag framework was to abolish the requirement of pre-defined dictionaries of chemical motifs, *e.g.*, functional groups, ring systems, synthesis artifacts such as amide bonds, etc. We pictured the available approaches that are based closely or loosely on chemical features (implementations of MACCS keys, FCFP or early graph reduction concepts by Gillet *et al*. and Barker *et al.*) as molecular graph coloring with a quite limited palette of colors. The number of colors, reflecting the number of different chemical epitopes or features, eventually determines the flexibility or fuzziness of the similarity assessment and thus impacts the method's capability of finding something new

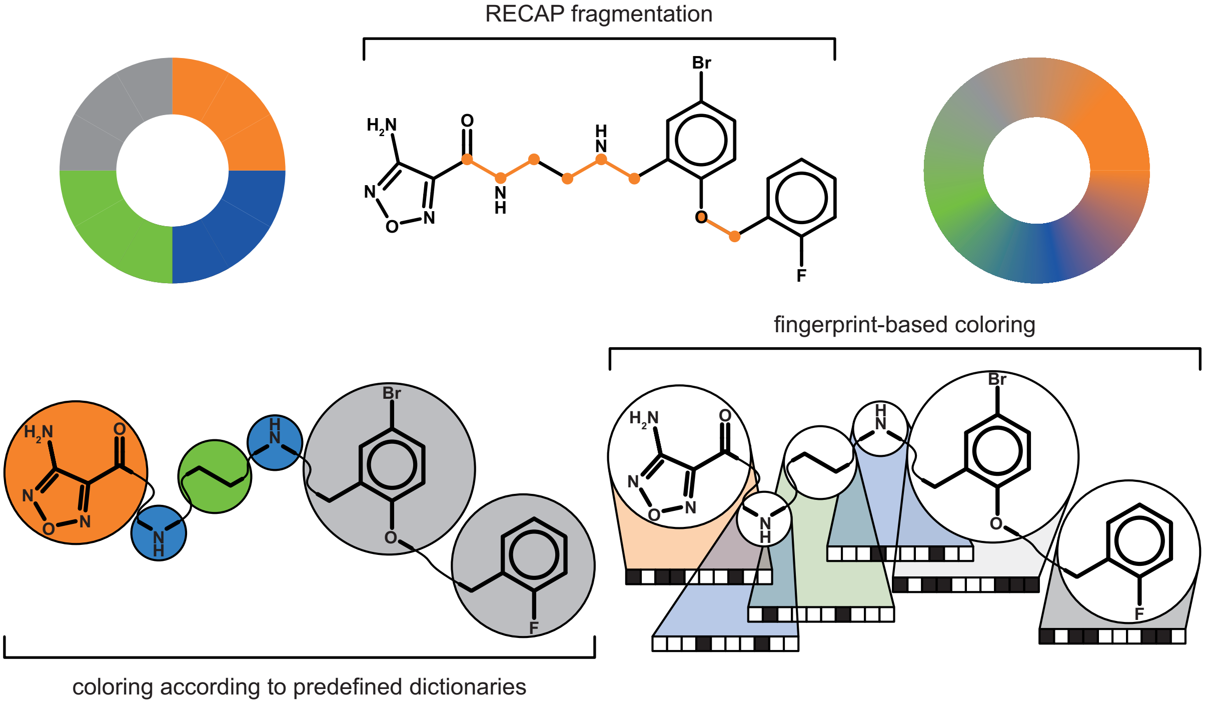

RECAP fragmentation
fingerprint-based coloring
coloring according to predefined dictionaries
Br
H_2N
O
N
H
F

Figure 1: RedFrag encodes the fragments originating from a molecule fragmented at the RECAP bonds (top, orange) with fingerprints (bottom right). In this case, the accessible color space is larger than the one spanned by a limited number of predefined dictionaries of chemical features (bottom left and illustrated by the 4-color cycle to the left and the continuous color palette to the right).

(**Figure 1**). RedFrag escapes this finite-color-space by exploiting the fragmentation of molecules as intuitive source of molecular features and chemical epitopes. The requirement of predefining these features is, within the context of RedFrag, transferred to a handful of breakable bond definitions: The definition of breakable bonds is of course not straightforward. Care must be taken if a reasonable and chemically sane fragmentation result is desired. However, bond definitions as defined by the RECAP or BRICS rules, reflect to some degree retrosynthetic routes and deliver sensible results. *In extremis*, a single bond definition can be applied in RedFrag:

[R]!@[!D1]

This SMARTS string shows the *golden* rule of fragmentation. The rule consists of two atom definitions and a bond definition. Enclosed in square brackets (green and purple) are atoms connected by any bond type except a ring bond (orange). This ensures the integrity of ring systems whose cleavage seems unreasonable from a chemist's point of view. This bond connects a ring atom (green) with an atom with two or more neighbors (purple). Acting upon a compound, this rule separates ring systems with their decoration still attached from linking moieties, eventually leading to an intuitive fragmentation pattern. The resulting fragments are then not encoded by colors from a limited color space (dictionaries) but described by a fingerprint. Molecular properties of each fragment such as its size, number of rotatable bonds, number of H-bond donor and acceptor functionalities and even further substructural motifs are – depending on the fingerprint in place – preserved and determine the color of the fragment. The color space defined in such a way is only limited by the resolution of the fingerprint and is in general much larger than the commonly used dictionaries of chemical motifs (**Figure 1**).

RedFrag introduces the infinite color space in terms of rule-based fragmentation and fingerprint encoding, thus allows for fuzzier similarity assessment. This translates to an increased variability of fragment-to-fragment comparisons and eventually allows for a more prominent motif exchange or replacement. Bioisosterism and scaffold hopping, for RedFrag, these concepts are within reach and without the need of conformational data.

We have tested RedFrag retrospectively on the MUV data sets. This choice ensures that the

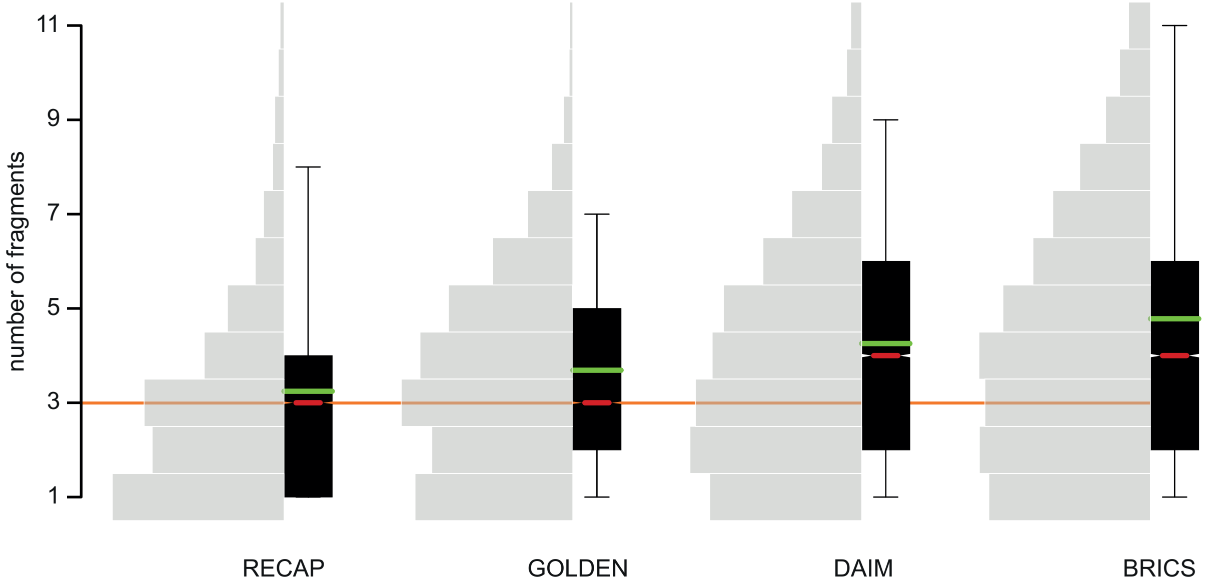


Figure 2: Distribution of molecular size in respect to the number of fragments in the reduced graph representations and the underlying fragmentation pattern. GOLDEN corresponds to the single-SMARTS rule mentioned in the main text. Orange line indicates a recommended minimum number of fragments per molecule. Red lines: median values; green lines: mean values of the distributions. Molecular database used for decomposition was the Chemical Component Dictionary (CCD): small molecules extracted from PDB entries (Westbrook *et. al.*, *Bioinformatics*, **2015**, *31 (8)*, 1274-1278).

observed performance is not distorted by trivial separations of actives and decoys, as often seen if self-compiled datasets, *e.g.*, from ChEMBL, are used for validation purposes. RedFrag performed at the same level as the other methods investigated on the MUV data set. Furthermore, a comparison between RedFrag and holistic, and highly optimized fingerprint-based similarity concepts, *e.g.*, ECFP, at their optimized parameter sets, show an on-par performance, overall. Interestingly, but not entirely unexpectedly, the performance of all investigated methods fluctuated throughout the validation set and the different target classes. It shows that the ultimate answer to the molecular similarity problem, as far as it exists at all, is not yet found. Going beyond retrospective analysis, we also applied RedFrag to predict binders for the aspartic protease endothiapepsin and tested our predictions *in vitro*.

RedFrag also has the potential to retrieve scaffolds distant from the query, *i.e.*, perform scaffold hops. This can be gleaned from **Figure 7**, **Part 1**, where the propensity of RedFrag to retrieve unusual scaffolds is comparable to other methods. This ability of RedFrag is also nicely demonstrated in the prospective screen, where the second and third most potent compounds are distant from the known ligand space of endothiapepsin, as evaluated by ECFP4 fingerprints. At the same time, the prospective screen also yielded a very potent molecule, which at 4 µM is one of the more potent binders of endothiapepsin. While it cannot be described as novel when compared to the molecules that were used as query, it is a clear improvement on the compound series published earlier.

In conclusion, we demonstrated that RedFrag is an intuitive, fast and unbiased algorithm for 2D molecule searches. The performance in both retrospective and prospective studies is in line with existing methods, but often yields different answers, adding it to the repertoire of suitable approaches for large-scale screenings. Most importantly, we abolished the need for predefining a color space without a loss in performance. This means that our algorithm is applicable to all sets of molecules without restrictions. RedFrag has only two main parameters, *i.e.*, choice of fragmentation pattern and fingerprint type: A well-chosen fragmentation pattern should decompose a query molecule into at least two, better into three fragments. This ensures that RedFrag topological replacement of fragments within the framework of a reduced graph can actually occur. **Figure 2** shows the distribution of the number of fragments

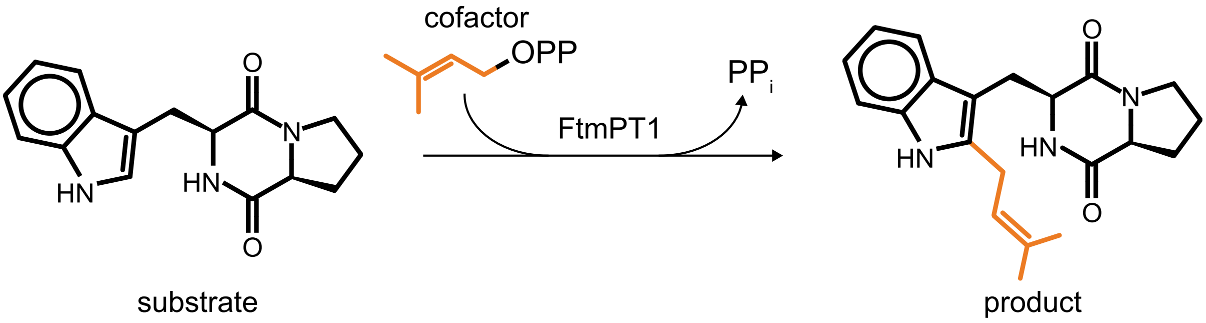


Figure 3: Attachment of a prenyl moiety (orange) by FtmPT1 prenyltransferase.

in respect to a fragmentation rule: Clearly, the fragmentation rule has a large impact on the complexity of generated reduced graphs and thus on the outcome of the similarity assessment. However, although also dependent on the molecular database used for screening, a general recommendation for DAIM and BRICS rules can be given, as they produce reduced graphs with a median size of four fragments. The choice of a fingerprint can be made based on a similar reasoning: the larger the number of fragments, the smaller they are. It is thus that an atomistic (ECFP) rather than a feature-based fingerprint (MACCS or FCFP) is the better choice.

2. Future of PrenDB: on the road to SAR-by-Enzyme

The **PrenDB** project involved a thorough literature search, the design and programming of novel data structures and the collection of experimental data. And yet, it is only the prelude to the main feature, which is SAR-by-Enzyme (**SARbyE**). The idea behind SARbyE is simple: Instead of using the standard medicinal-chemical diversification repertoire, *e.g.*, adding a methyl group, capping hydroxy groups or attaching phenyl groups to the hit compound, why not let enzymes do the job? Enzymes catalyze chemical reactions under ambient conditions, at room

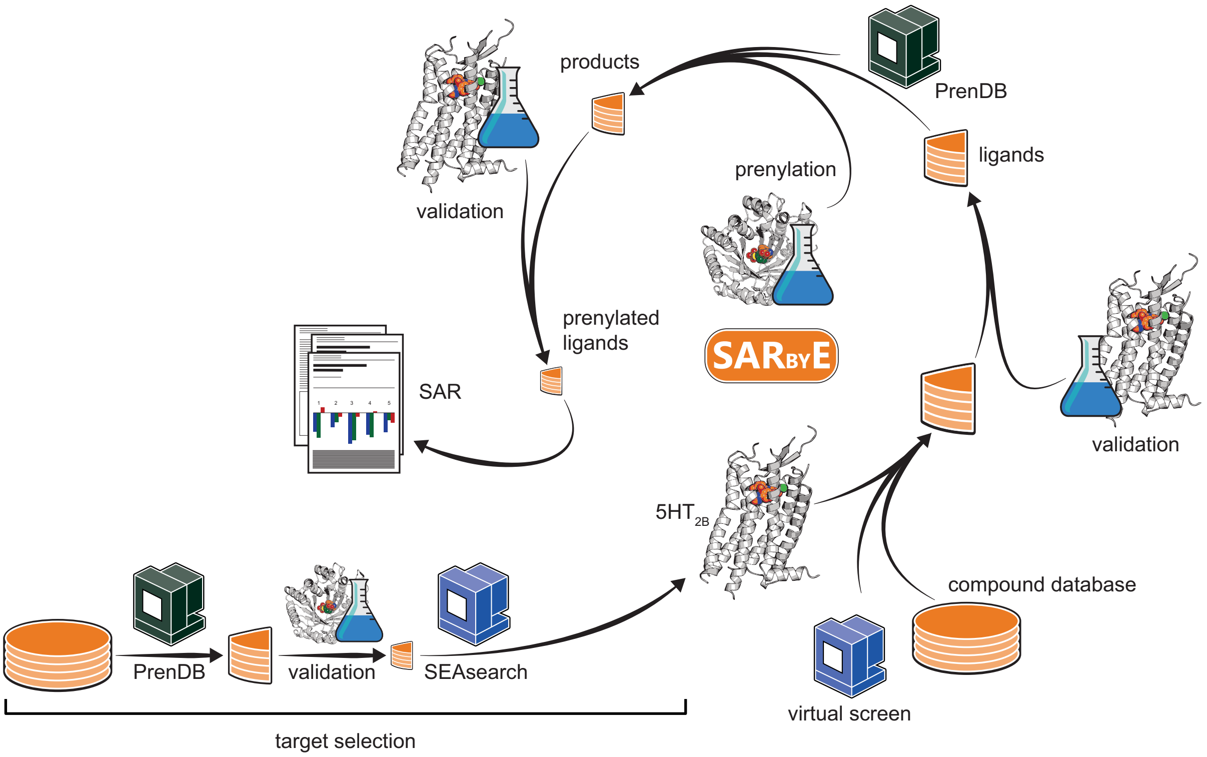


Figure 4: Concept of SAR-by-Enzyme. With PrenDB, the substrate space of prenyltransferases has been expanded and used by the SEAsearch algorithm to predict a suitable target. The iteration of *in silico* predictions and experimental validation resembles the SARbyE cycle, which delivers the SAR data for prenylated – diversified compounds.

temperature, in a buffered aqueous solution. They convert molecules under stringent stereo- and regioselective control and fulfil the criteria of green chemistry, which becomes ever more important and acknowledged.

Biocatalysis – the actual use of enzymes to accelerate chemical reactions in an academic or industrial setting – is not new but it gains popularity and applicability since the dawn of modern enzyme engineering techniques.

In SARbyE, we want to exploit a promiscuous enzyme family, the dimethylallyltryptophan synthase family (DMATS), in order to establish an alternative and attractive route to compound diversification and optimization. Prenyltransferases, an extensively studied enzyme class and member of the DMATS family, catalyze such an attractive reaction (**Figure 3**): Prenyl moieties are small apolar, flexible groups that by increasing the lipophilicity of the compound and the lack of the necessity of directionality, increase the binding affinity of the compound and open additional routes for further diversification.

In order to use prenyltransferases as a biocatalytic toolkit in the prospects envisioned in SARbyE, the development of PrenDB was a necessary step: PrenDB is a catalog of prenyltransferase reaction which were extracted from primary literature and compiled into a browsable format. Moreover, it is capable of predicting the prenylability of compounds by investigating the compound's chemical epitopes and comparing them with stored repitopes from known substrates. Within the larger picture of SARbyE (**Figure 4**) PrenDB plays an important role: After the initial virtual screening and the experimental validation of the putative binders to the selected target, predictive routines of PrenDB are responsible to evaluate the prenylability of the hits. Putative substrates are then subjected to actual prenylation and later their binding towards the initial target is revaluated and the SAR established.

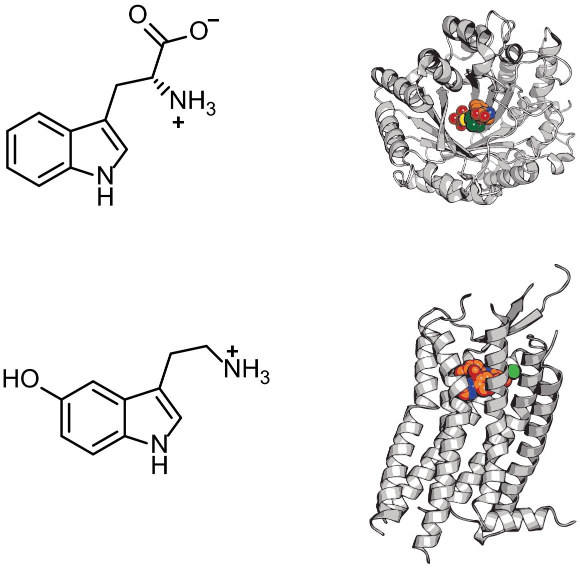


Figure 5: Comparison of tryptophan, natural substrate for several prenyltransferases (top row), and 5-hydroxytryptamine (5-HT or serotonin), the natural ligand of the serotonin receptor 5-HT$_{2B}$ (bottom row).

In PrenDB we demonstrated the power of systematically organizing and analyzing experimental enzymatic data by means of chemoinformatic methods. It is a comprehensive repository of the existing knowledge about prenyltransferase reactions. With our algorithmic determination of what we called repitopes (reactive epitopes) we were able to predict novel substrates. In a symbiotic manner, we used these predictions to identify a potential test target for the SARbyE concept (**Figure 4**, bottom left). We identified the serotonin receptor 5-HT$_{2B}$ as promising target with a ligand space comparable to the substrate space of prenyltransferases (**Figure 5**).

However, the repitopes stored in PrenDB are not yet accurate enough in all cases to precisely predict the prenylability. This shortcoming is presumably correlated with the comparatively small number of instances in the database. Although the existing body of literature clearly represents a considerable experimental effort, chemistry and the biochemical reactivity of enzymes are so diverse that even higher numbers of substrate-enzyme-product triplets would be necessary to obtain more complete repitopes that also account for the different reactivity of certain substructures.

At the same time, a database such as PrenDB can provide excellent help in determining which reactions and substrates would be worthwhile to test next. On a basic level, one could simply be guided by the number of reactions already described for each enzyme and focus on the underrepresented ones. But also, more sophisticated approaches can be envisioned:

enzyme phylogenetic trees could be based not on amino acid sequence, but on substrate similarity. Further exploration would thus focus on filling in the *missing links*. Ultimately, such strategies might merge with machine learning approaches, where the algorithm itself would suggest which enzyme-substrate pairs to test next based on the maximum information gain of each investigation.

Lastly, it must be emphasized that the concept of repitopes and their fragment-based determination can easily be extended to other enzymatic reactions. The automatic processing of potentially large numbers of reactions and the concomitant conversion into the reaction principles will lead to facile systematizations and gain of knowledge from the analyses of the emerging data.

3. How can we understand regioselectivity of chemoenzymatic reactions?

In our vision of SARbyE, PrenDB, in addition to its storage and retrieval functions of prenyltransferase reactions, is a knowledge-based prediction backend designed to derive reactive epitopes (repitopes) from reaction data and use them in a substructure-based search for novel substrates for prenyltransferases. Its predictive power relies on the amount of data stored within and the novelty of the predictions depends on the chemical diversity of the compounds prenylated so far. It seems thus intuitive and important to expand the reaction space of prenyltransferases by experimentally testing a variety of compounds and chemical scaffolds. In addition to prenyl acceptors, *e.g.*, L-tryptophan and brevianamide F (**E1**, **Figure 1**, **Part 2**), a diversification in the reaction space can be gained from unnatural prenyl donors. In **Part 3** of this thesis, we presented five prenyltransferases and investigated their acceptance of not only dimethylallyl pyrophosphate (DMAPP) but additional three unnatural prenyl donors: Methylallyl, 2-pentenyl and benzyl pyrophosphate, MAPP, EAPP and BENZYLPP, respectively. L-tryptophan was used as the prenyl acceptor.

We observed a total of eleven prenylation products with fluctuating conversion rates throughout the deck of prenyltransferases (**Scheme 2**, **Part 3**). The multitude of different regioselectivities and yields raised the question how the enzymatic prenyl transfer is controlled at the atomistic level and whether the unique constellation of the active site residues of a prenyltransferase, the electro-sterical properties of the prenyl donors and the presumably comparable reactivity of the prenyl acceptor can be put into agreement with the observed experimental data.

Knowledge gained from the mechanics of the prenylation reaction can be transferred into mutant models that allow for the engineering of prenyltransferases into more efficient, more promiscuous and more widely applicable molecular diversification machines as envisioned by the SARbyE concept. Thus, we decided to use molecular dynamics simulations in order to pursue the interactions of enzyme, prenyl donor and acceptor with each other and their relative orientation to each other.

By using the program MODELLER, a structural model was constructed for 5-DMATS from *A. clavatus* and used for docking and molecular dynamics studies with DMAPP, MAPP, EAPP and BENZYLPP, leading to a distance-based explanation of their observed reaction preferences (**Figure 3**, **Part 3**). Furthermore, we were able to extract system properties such as root-mean-square fluctuations (RMSF) of the atom positions of L-tryptophan induced by the presence of the prenyl donors (**Figure 5**, top row): It shows that benzyl pyrophosphate deteriorates the position of the prenyl acceptor more strongly than other donors (blue line) which is in accordance to the overall instability of the simulation conducted with benzyl pyrophosphate (**Figure 6**, middle row) and in accordance with the experimental data reported in **Scheme 2**, **Part 3**. However, it is apparent that this stochastic approach is not accurate enough to explain the observed data universally or to be used prospectively, *i.e.* for prediction of reactivity and regioselectivity of a prenyltransferase and a prenyl donor/acceptor pair. In order to elucidate the different aspects of prenylation reactions, three modifications of, or extensions to, the molecular dynamics simulations protocol used in our study can be made: i) Under the assumption that reactivity of a given prenyl acceptor correlates with its binding affinity to the enzyme-prenyl-donor complex, free energy perturbation (FEP) calculations can be conducted. There, a prenyl acceptor, *e.g.*, DMAPP, is alchemically transformed during the simulation into another one, *e.g.*, MAPP. This transition of chemical species, which is an application of the more general thermodynamic integration approach (TI), allows for the estimation of relative free-energy of binding (ΔΔG values). From these values, it can be judged whether a distinct combination of prenyl donor and acceptor (and possible mutations in the active site) are favorable in terms of energy and thus more reactive. Of note, FEP, driven by classical molecular mechanics, is not able to render the actual enzymatic reaction. Bond cleavage and formation, thus reallocation of electrons can accurately be captured by quantum mechanics, only. The combined approach, quantum mechanics/molecular mechanics molecular dynamics (QM/MM MD), is resource-demanding and rarely applied. ii) The reactivity of chemical matter in an enzymatic setup is not only controlled by thermodynamics, *e.g.*, interaction energy and entropic effects, but also by kinetics. A substrate has to associate rapidly with the enzyme, stay long enough in active site to be transformed into the product and dissociate fast in order to free the active site for the next substrate molecule. Association and dissociation rates, k_{on} and k_{off}, respectively, cannot be captured by methods such as FEP. In order to estimate these values by molecular dynamics the association of a compound to the target – its translation from bulk solvent to the interior of the targets binding or active site – has to be simulated. Albeit acceleration techniques such as hyperdynamics, parallel replica dynamics, temperature accelerated dynamics and steered molecular dynamics, estimation of k_{on} and k_{off} is very demanding in computational time and not yet applicable for a large number compounds in a predictive screening scenario.

4. Synergistic triad of molecular similarity, substructure search and docking

Throughout this thesis and the studies described therein, three methodological pillars of computer

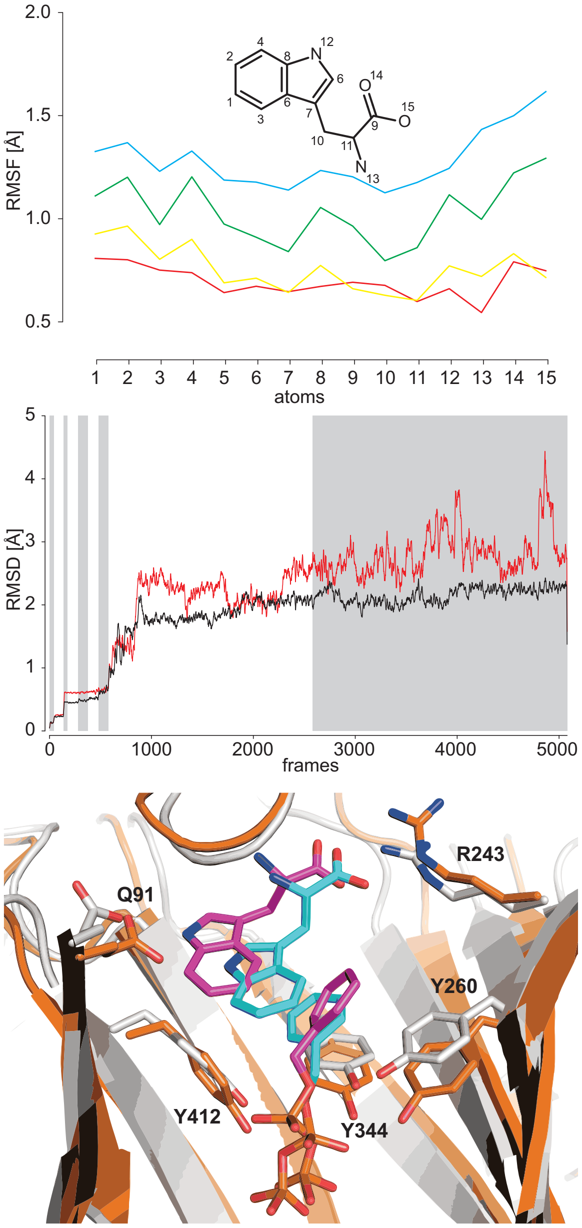


Figure 6: Analysis of molecular dynamics simulation data of prenyltransferase complexes. Top row shows root-mean-square fluctuations (RMSF) of L-tryptophan (TRP) in complex with different donor molecules. Blue: TRP and benzyl pyrophosphate; green: TRP and 2-pentenyl pyrophosphate; yellow: TRP and dimethylallyl pyrophosphate; red: TRP and methylallyl pyrophosphate. Middle row shows the root-mean-square deviation (RMSD) of the coordinates of the TRP molecule (red) and benzyl pyrophosphate (black) during each step of the simulation indicated by white and gray background, respectively. Bottom row shows complexes of the prenyltransferase 5-DMATS with L-tryptophan and benzyl pyrophosphate. Colored in orange and cyan is the equilibrated complex before the productive molecular dynamics run. Colored in white and magenta is the complex after 5 ns productive simulation run.

aided drug discovery and design have been in focus: i) Fingerprint-based similarity, ii) substructure-based searches and iii) molecular docking. These concepts were used as standalone tools or were combined into a hierarchical screening cascade in order to fetch promising compounds from vast molecular databases – to find the needle in the haystack. Moreover, the conceptual basis of molecular similarity and substructure-based search (graph isomorphism) was the inspiring starting point for the development of novel tools and methods. In either way, end-user application of, *e.g.*, commercial docking tools and highly optimized fingerprints, or the design of new algorithms and the combination of robust and well-understood concepts, the synergistic relationship of these tools, ideas, methods is, although often unintentionally, ubiquitous.

One example for this hidden synergy – contextually also complementarity or orthogonality – is given by the screening study described in **Part 4** of this thesis. There, a novel class of structural motifs, a novel scaffold, was exploited as a seed for the expansion of the ligand space of the β$_2$-adrenergic receptor (β$_2$AR). In particular, the screening conducted in this study was a combination of a *horizontal* and *vertical* one (*e.g.*, **Figure 7**). The fingerprint-based similarity search on the one hand and the substructure-based search on the other efficiently reduced the compound database from millions to several hundred thousand partly overlapping compound sets. Indeed, the partial overlap is the key here, as both methods deliver to some extent orthogonal answers, *i.e.*, molecules considered similar to the query but also unique for the given retrieval method. This orthogonality or complementarity seems irritating at the first glance because both methods work under the

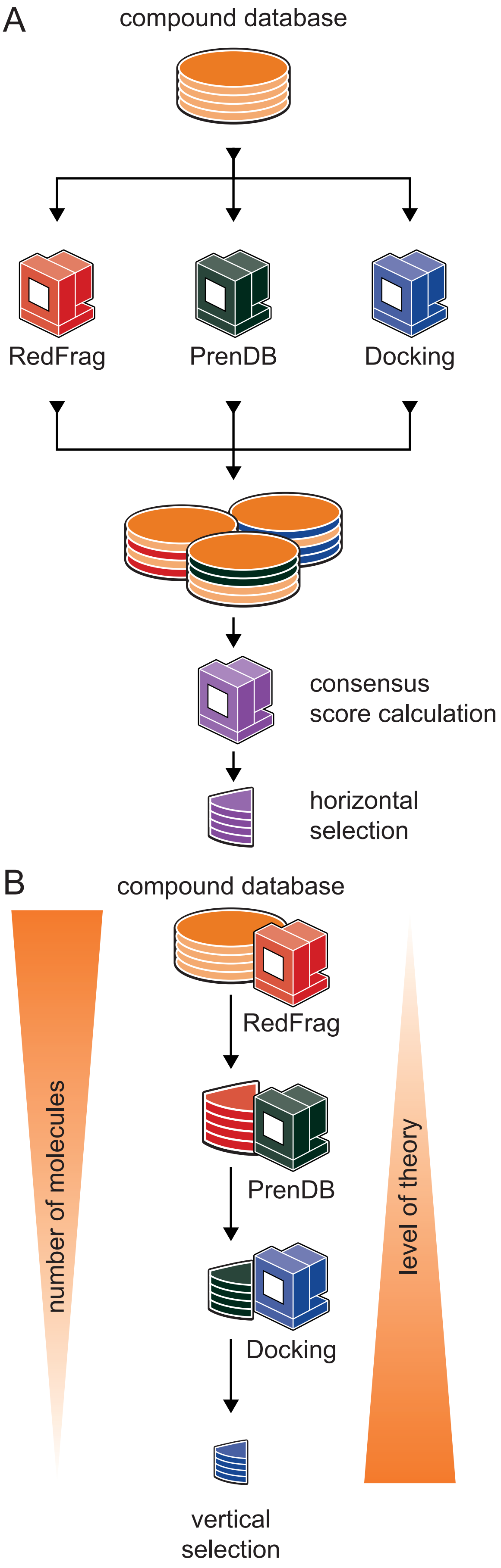


Figure 7: A Horizontal or orthogonal screening approach. A compound database is ranked or scored independently by multiple screening methods, *e.g.*, RedFrag's 2D-similarity, PrenDB substructure-based search and molecular docking. The output is then combined based on the consensus score: A mathematical formula incorporating the individual scores and/or weighting coefficients giving the horizontal selection. **B** Vertical or hierarchical screening cascade starting with fast but less accurate methods (fingerprint-based or pharmacophore searches) and concluding with slow but most accurate methods (docking, molecular dynamics).

premise to retrieve entities with similar properties (*e.g.*, similar bioactivity). Hence, intuitively, the overlap of compounds across different methods should be large and method-exclusive compounds the exception. However, the reality is just the contrary of the human's intuitive mind and the discrepancy between the retrieval is even more pronounced for structure-based methods such as docking, where the score – the likelihood of the generated geometry to be the correct bioactive one – correlates only poorly or not all across docking algorithms and tools. How to use the output of a horizontal screening then? A widely-applied approach is the calculation of a consensus score. A score that reflects the quality of a compound estimated by each of the used methods. It emphasizes compounds that were highly ranked, scored or favored by multiple methods and deprioritizes compounds that were well-scored only occasionally. In such a way, it seems possible to extract molecules with an enhanced confidence being, *e.g.*, active.

Alternatively, the vertical screening approach seems to be more reasonable: It is a cascade of the sequential invocation of screening tools where the output of a method, *e.g.*, most similar compounds as determined by fingerprint-based similarity, is subjected as input to the next method, *e.g.*, pharmacophore search. This approach reduces the number of compounds that has to be screened from method to method and thus allows for the usage of more time-demanding tools and accurate algorithms further down in the screening cascade. At the end of this hierarchical screening a vertical set of compounds emerges that caries the best estimates of binding affinity or related quantity of the whole set of invoked methods. Similar to the horizontal set, these compounds show an enhanced confidence in their estimated property as they were repeatedly well-scored from the top to the bottom of the screening cascade. However, there is major difference: The vertical set of compounds is in its composition strongly biased by the first method used at the top of the cascade. Thus, in dependence of the method in place, the outcome of the vertical screen, more precise its diversity, is limited to the capability of detecting diverse chemical matter of the method used. It is thus crucial to design the initial screening step wisely, *e.g.*, allowing for scaffold-hoping-enabled methods early in the cascade and still maintaining the calculation speed. **RedFrag** screening, for 2D-based, and ROCS analysis or pharmacophore search for 3D-based methods are reasonable choices as demonstrated in **Part 2** (**PrenDB**) of this thesis.

Given the example of the β$_2$AR-screening presented in **Part 4**, the combination of horizontal and vertical screening and, eventually, their beneficial combination is well described: Fingerprint-based similarity and substructure-based search are concluded by a consensus calculation, being in this case a simple unification of the two compound sets. The consensus compound set was than subjected to a docking algorithm, which narrows down the number of molecules further. The second and last step in the vertical approach was the visual inspection of the generated geometries. This introduced the human intuition and knowledge into the screening and culminated the estimation of the likelihood of the compounds to behave similarly as the query compounds used at the beginning of the screening.

Grow old along with me!
The best is yet to be,
The last of life, for which the first was made:
Our times are in His hand
Who saith “A whole I planned,
Youth shows but half; trust God: see all, nor be afraid!”

from **Rabbi Ben Ezra** by **Robert Browning**, † 1889.

Backmatter

Eidesstattliche Erklärung

Gemäß §10 Absatz (1) b der Promotionsordnung vom 22.04.2009 versichere ich, dass ich meine Dissertation mit dem Titel

"Development and prospective application of chemoinformatic tools to explore new ligand chemistry and protein biology"

selbstständig ohne unerlaubte Hilfe angefertigt und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen bedient habe. Alle vollständig oder sinngemäß übernommenen Passagen sind Zitate und als solche gekennzeichnet.

Gemäß §10 Absatz (1) c der Promotionsordnung vom 22.04.2009 erkläre ich, das die Dissertation in der jetzigen oder einer ähnlichen Form noch bei keiner anderen Hochschule eingereicht wurde und noch keinem sonstigen Prüfungszweck gedient hat.

Marburg, den 19.05.2017

____
(Jakub Gunera)

CV

Aus Gründen des Persönlichkeitsschutzes wird von der elektronischen Veröffentlichung des Lebenslaufes abgeshen.

Aus Gründen des Persönlichkeitsschutzes wird von der elektronischen Veröffentlichung des Lebenslaufes abgeshen.

Aus Gründen des Persönlichkeitsschutzes wird von der elektronischen Veröffentlichung des Lebenslaufes abgeshen.

Aus Gründen des Persönlichkeitsschutzes wird von der elektronischen Veröffentlichung des Lebenslaufes abgeshen.

Supporting Information

Fragment-based similarity searching with infinite color space

Jakub Gunera[1] and Peter Kolb[1]

[1]Department of Pharmaceutical Chemistry, Philipps-University, Marburg, Hesse, 35032, Germany

1. Maximum Unbiased Validation data sets (MUV)

The 17 activity classes were designed with aim to minimize both artificial enrichment as well as analogue bias. Further filters were employed to remove aggregators and unspecific binders. **Table S1** categorizes the activity classes by the target class and the mode of interaction of the underlying active molecules.

Table S1: Composition of the Maximum Unbiased Validation data sets (MUV).

Tag	Target	Mode of interaction	Target class
S1P1R	S1P1 receptor	agonists	GPCR
PKA	PKA	inhibitors	kinase
SF1I	SF1 receptor	inhibitors	nuclear receptor
RK2	Rho-Kinase2	inhibitors	kinase
HIV	HIV RT-RNase	inhibitors	ribonuclease
EPHA4	EPH receptor A4	inhibitors	protein-tyrosine kinase
SF1A	SF1 receptor	agonists	nuclear receptor
HSP90	HSP 90	inhibitors	chaperone
ERAI	ER-α-coactivator binding	inhibitors	protein-protein interaction
ERBI	ER-β-coactivator binding	inhibitors	protein-protein interaction
ERAP	ER-α-coactivator binding	potentiators	protein-protein interaction
FAK	FAK	inhibitors	kinase
CATG	Cathepsin G	inhibitors	protease
FXIA	FXIa	inhibitors	protease
FXIIA	FXIIa	inhibitors	protease
D1R	D1 receptor	allosteric modulators	GPCR
M1R	M1 receptor	allosteric inhibitors	GPCR

2. Comparison of fingerprint metrics

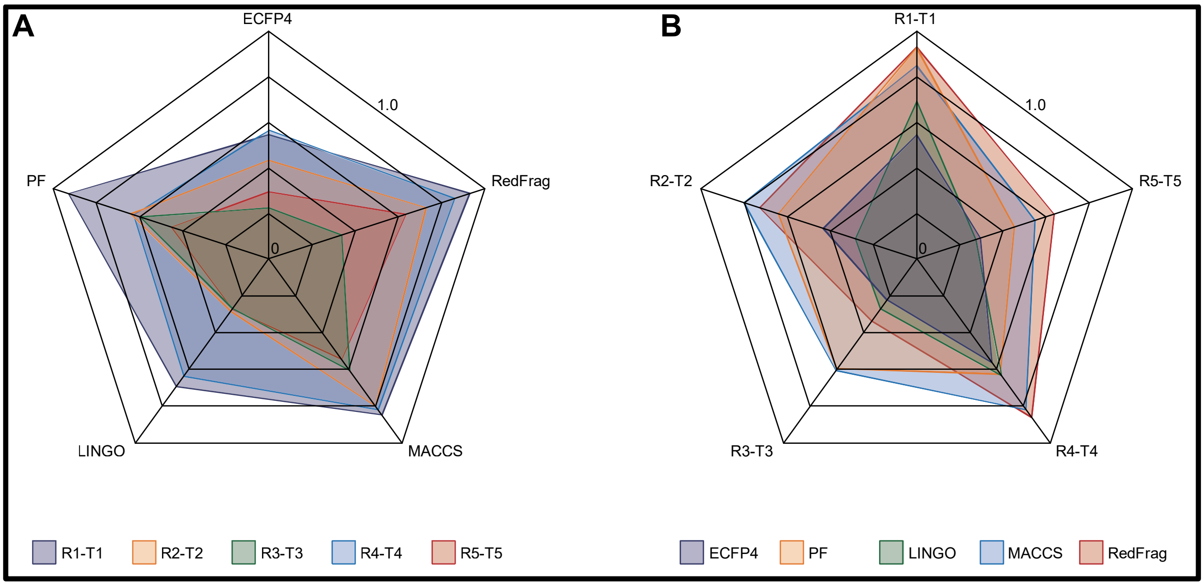


Figure S1: Radar plot representations of similarity distances calculated with five metrics (ECFP4, Pharmacophore fingerprint [PF], MACCS Keys, LINGO and RedFrag) for five molecule pairs (R1-T1 – R4-T5). **A** Molecule-pair-wise projection of similarity distances. **B** Metric-wise projection of similarity distances.

3. Properties of RFGraphs

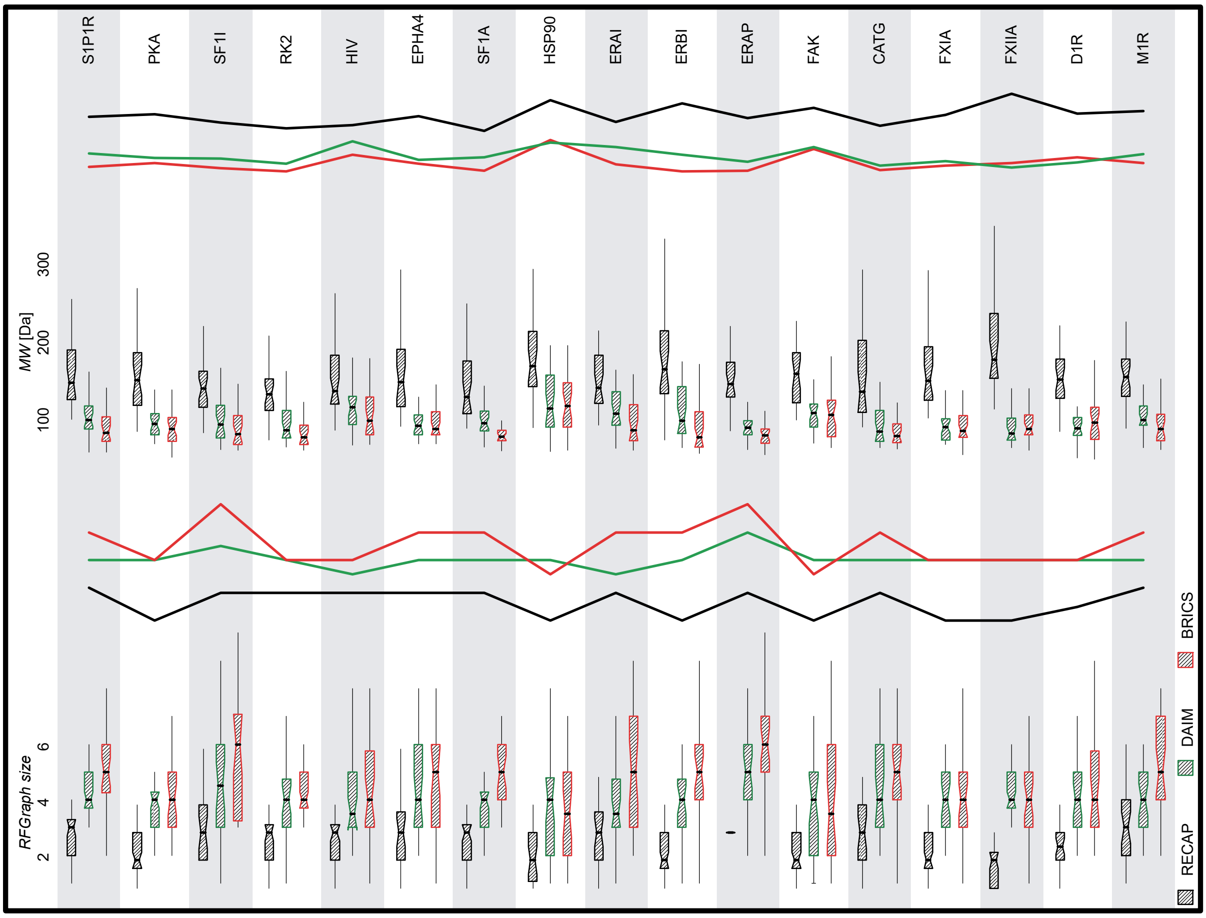


Figure S2: RFGraph size distributions and average fragment molecular weights for each activity class according to RECAP, DAIM and BRICS fragmentation rules. Solid lines emphasize the trend of median values of the underlying distributions.

4. Performance of fingerprints without fragmentation algorithm

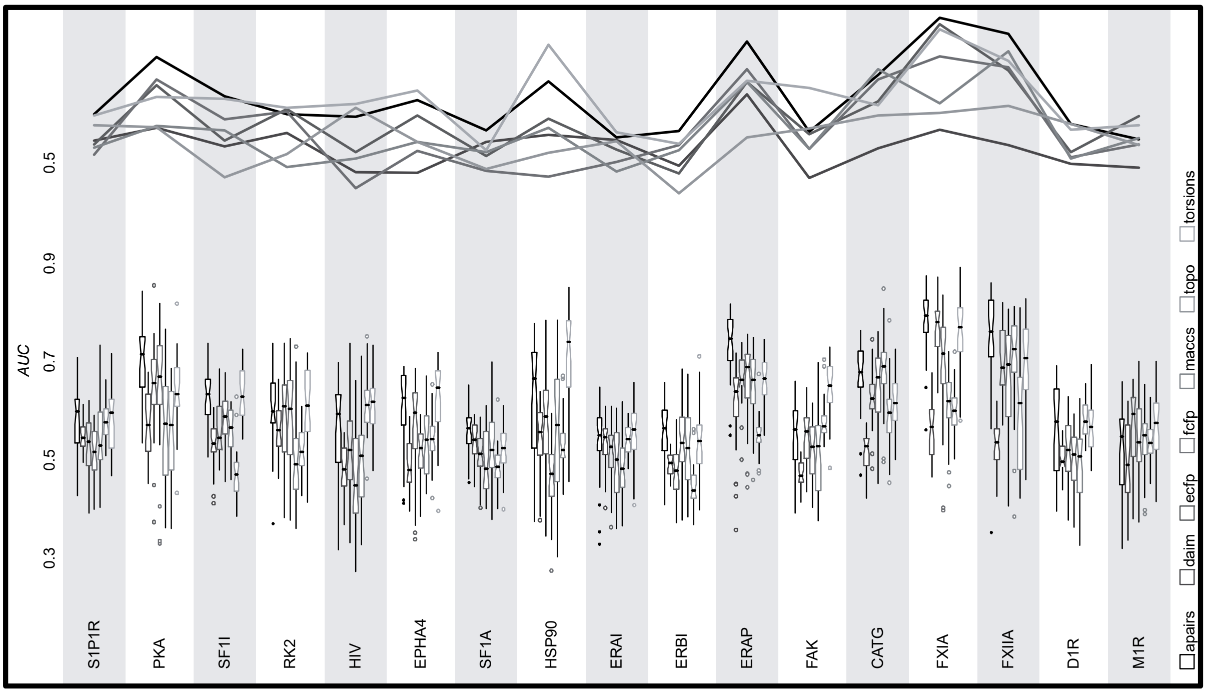


Figure S3: Shown are distributions of AUC values for seven fingerprints and for each activity class as box-plot representations. Solid curves emphasize the trend of the medians.

5. ECFP4 similarity of query molecules and virtual hits

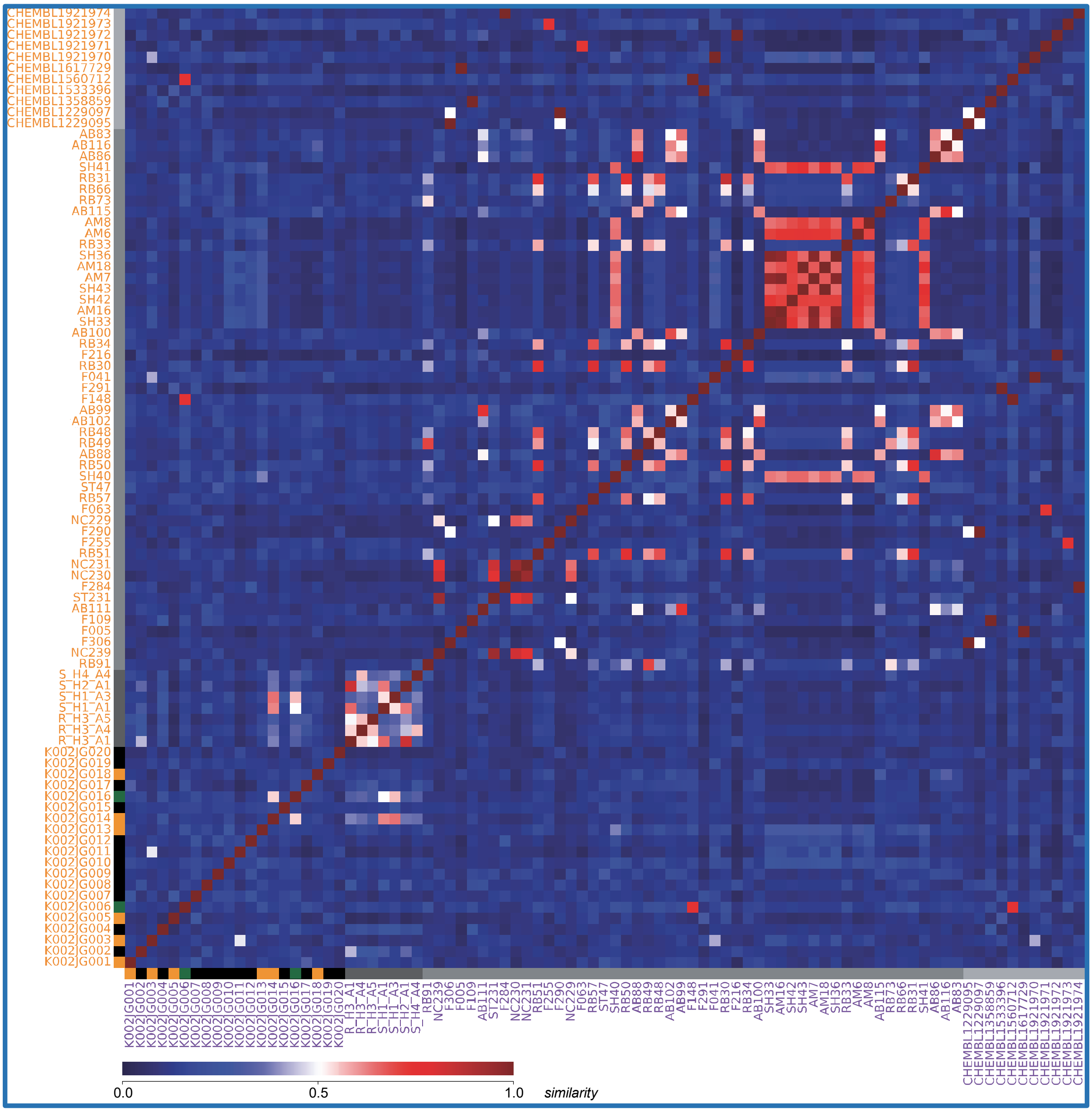


Figure S4: Similarity matrix of selected virtual hits (black bar), query compounds from a recently published ETP inhibitor design study (45), from ref. (46) and extracted from the ChEMBL database (44) (in this order of grayish bars). Yellow squares indicate experimentally validated hits (Table 2). Green squares indicate virtual hits with a high Tanimoto similarity (close analogues or homologues) to a known ETP inhibitor but without biological activity.

6. Scaffold similarity of query molecules and virtual hits

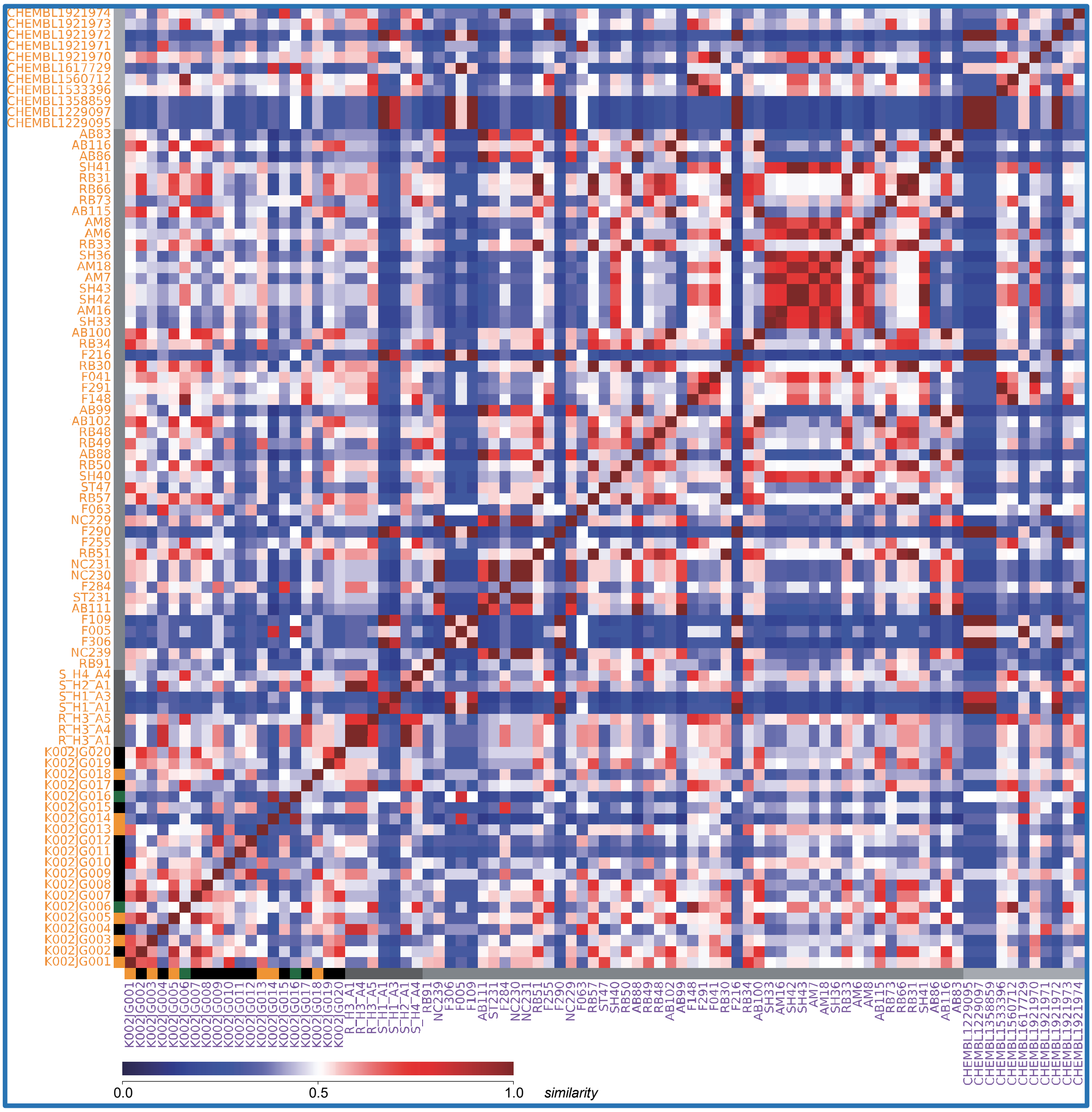


Figure S5: Scaffold similarity matrix of selected virtual hits (black bar), query compounds from a recently published ETP inhibitor design study (45), from ref. (46) and extracted from the ChEMBL database (44) (in this order of grayish bars). Yellow squares indicate experimentally validated hits (Table 2). Green squares indicate virtual hits with a high Tanimoto similarity (close analogues or homologues) to a known ETP inhibitor but without biological activity.

7. RedFrag similarity of query molecules and virtual hits

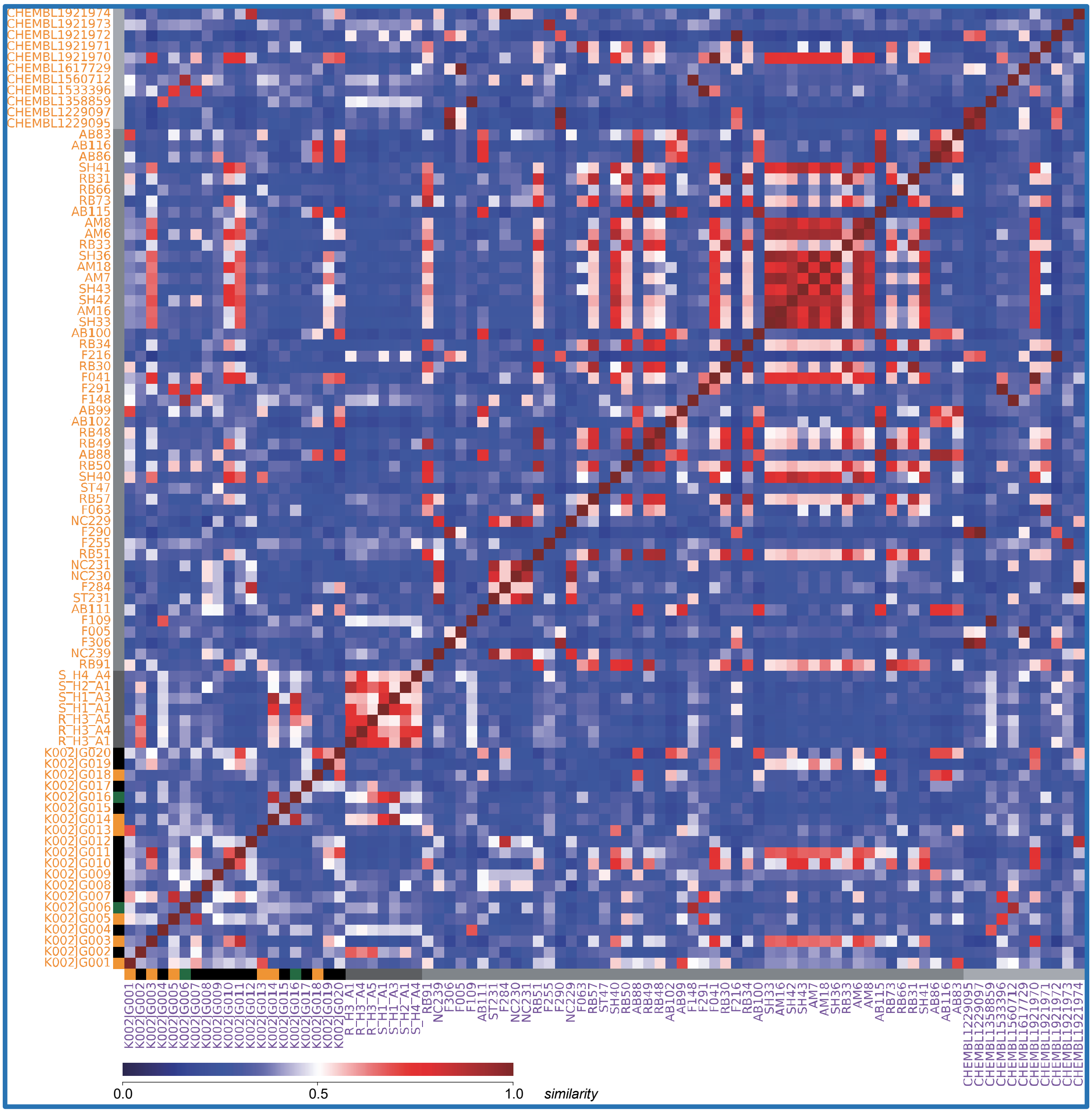


Figure S6: RedFrag similarity matrix of selected virtual hits (black bar), query compounds from a recently published ETP inhibitor design study (45), from ref. (46) and extracted from the ChEMBL database (44) (in this order of grayish bars). Yellow squares indicate experimentally validated hits (Table 2). Green squares indicate virtual hits with a high Tanimoto similarity (close analogues or homologues) to a known ETP inhibitor but without biological activity. The matrix was generated with RECAP|maccs|0.1|0.5-3.5 parameter set.

8. Endothiapepsin bioactivity assay

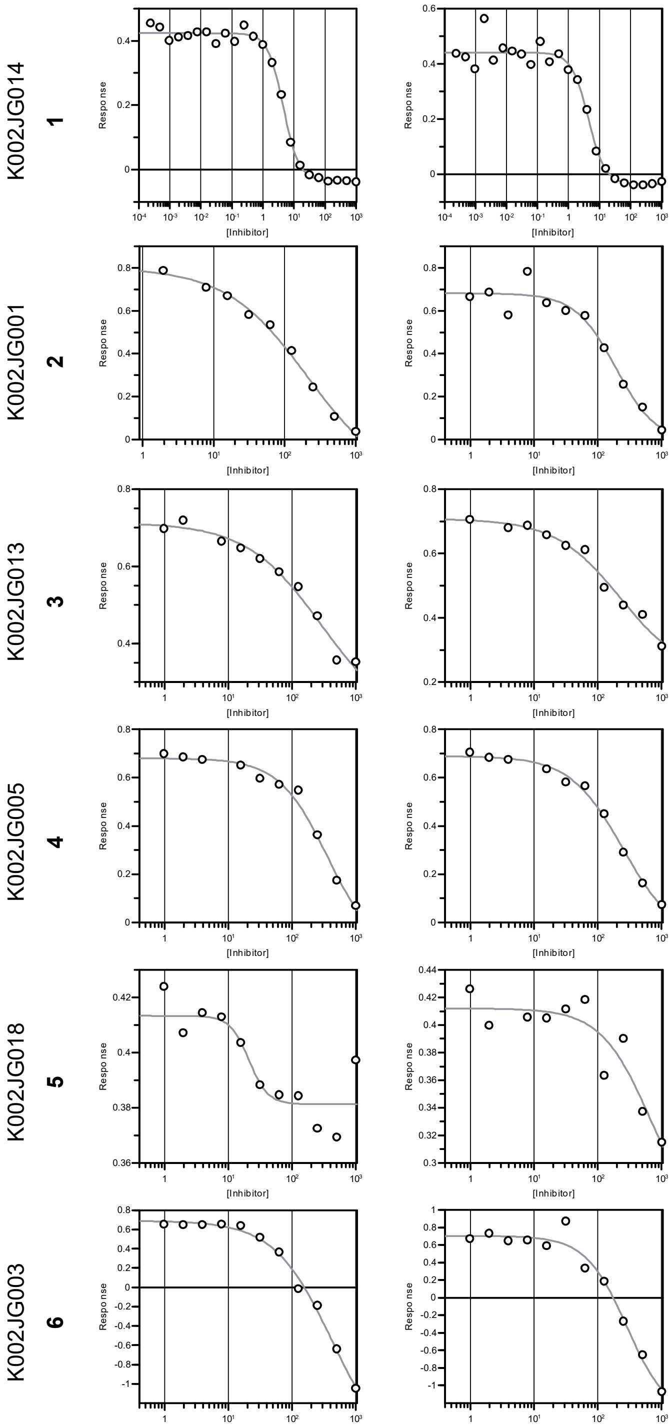


Figure S7: Dose response curves for hits identified in this study.

9. Retrospective MUV screening: Average median AUC values

9.1. RECAP

Table S2: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and Atom Pairs.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5630	0.4506	0.3034	0.1450	0.0981	0.0596	0.0505	0.0467	0.0467
3.0-1.0	0.5623	0.4502	0.3024	0.1451	0.0981	0.0596	0.0505	0.0467	0.0467
2.5-1.5	0.5620	0.4509	0.3024	0.1449	0.0980	0.0595	0.0505	0.0467	0.0467
2.0-2.0	0.5611	0.4506	0.3015	0.1448	0.0978	0.0595	0.0505	0.0467	0.0467
1.5-2.5	0.5561	0.4496	0.2998	0.1448	0.0978	0.0595	0.0505	0.0467	0.0467
1.0-3.0	0.5499	0.4465	0.2985	0.1447	0.0978	0.0596	0.0505	0.0467	0.0467
0.5-3.5	0.5394	0.4419	0.2977	0.1445	0.0977	0.0595	0.0505	0.0467	0.0467

Table S3: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and DAIM fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5325	0.5342	0.5395	0.5422	0.5474	0.5477	0.5392	0.5061	0.4108
3.0-1.0	0.5327	0.5346	0.5404	0.5425	0.5475	0.5470	0.5383	0.5059	0.4108
2.5-1.5	0.5354	0.5371	0.5416	0.5468	0.5503	0.5468	0.5379	0.5059	0.4104
2.0-2.0	0.5449	0.5462	0.5473	0.5534	0.5539	0.5510	0.5379	0.5063	0.4098
1.5-2.5	0.5579	0.5589	0.5587	0.5594	0.5583	0.5506	0.5377	0.5053	0.4095
1.0-3.0	0.5641	0.5628	0.5604	0.5603	0.5537	0.5456	0.5355	0.5067	0.4096
0.5-3.5	0.5663	0.5641	0.5598	0.5531	0.5456	0.5341	0.5276	0.5002	0.4077

Table S4: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and ECFP fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5353	0.3517	0.1516	0.0903	0.0655	0.0477	0.0467	0.0467	0.0467
3.0-1.0	0.5374	0.3500	0.1517	0.0902	0.0654	0.0477	0.0467	0.0467	0.0467
2.5-1.5	0.5368	0.3491	0.1519	0.0902	0.0654	0.0477	0.0467	0.0467	0.0467
2.0-2.0	0.5363	0.3476	0.1519	0.0902	0.0654	0.0477	0.0467	0.0467	0.0467
1.5-2.5	0.5344	0.3461	0.1520	0.0902	0.0654	0.0477	0.0467	0.0467	0.0467
1.0-3.0	0.5310	0.3438	0.1518	0.0903	0.0654	0.0477	0.0467	0.0467	0.0467
0.5-3.5	0.5220	0.3416	0.1518	0.0903	0.0654	0.0477	0.0467	0.0467	0.0467

Table S5: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and FCFP fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5520	0.5534	0.4383	0.3025	0.2069	0.1082	0.0739	0.0607	0.0559
3.0-1.0	0.5519	0.5534	0.4363	0.3022	0.2068	0.1082	0.0740	0.0607	0.0559
2.5-1.5	0.5523	0.5499	0.4346	0.3015	0.2067	0.1081	0.0739	0.0606	0.0559
2.0-2.0	0.5528	0.5462	0.4325	0.3013	0.2066	0.1081	0.0739	0.0606	0.0559
1.5-2.5	0.5521	0.5402	0.4308	0.3007	0.2068	0.1080	0.0739	0.0606	0.0559
1.0-3.0	0.5522	0.5329	0.4293	0.2995	0.2068	0.1080	0.0739	0.0606	0.0559
0.5-3.5	0.5472	0.5235	0.4255	0.2979	0.2068	0.1078	0.0737	0.0606	0.0559

Table S6: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and MACCS keys fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5478	0.5705	0.5584	0.4650	0.3282	0.1958	0.1181	0.0719	0.0553
3.0-1.0	0.5511	0.5732	0.5594	0.4607	0.3267	0.1954	0.1178	0.0717	0.0553
2.5-1.5	0.5592	0.5782	0.5591	0.4582	0.3264	0.1952	0.1175	0.0717	0.0552
2.0-2.0	0.5685	0.5820	0.5571	0.4560	0.3261	0.1949	0.1175	0.0717	0.0552
1.5-2.5	0.5758	0.5817	0.5543	0.4557	0.3255	0.1945	0.1174	0.0717	0.0552
1.0-3.0	0.5846	0.5802	0.5526	0.4550	0.3247	0.1944	0.1172	0.0715	0.0552
0.5-3.5	0.5864	0.5772	0.5489	0.4537	0.3239	0.1942	0.1172	0.0715	0.0552

Table S7: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and Topological fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5538	0.5599	0.5579	0.4864	0.1846	0.1008	0.0657	0.0522	0.0476
3.0-1.0	0.5548	0.5590	0.5572	0.4857	0.1840	0.1005	0.0657	0.0522	0.0476
2.5-1.5	0.5577	0.5598	0.5557	0.4827	0.1837	0.1005	0.0657	0.0522	0.0476
2.0-2.0	0.5618	0.5595	0.5541	0.4796	0.1833	0.1006	0.0657	0.0522	0.0476
1.5-2.5	0.5624	0.5598	0.5510	0.4768	0.1831	0.1006	0.0657	0.0522	0.0476
1.0-3.0	0.5563	0.5527	0.5435	0.4721	0.1826	0.1005	0.0648	0.0522	0.0476
0.5-3.5	0.5462	0.5375	0.5255	0.4644	0.1822	0.1004	0.0655	0.0522	0.0476

Table S8: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and Topological torsions fingerprint .

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.4665	0.2318	0.1304	0.0782	0.0607	0.0487	0.0339	0.0279	0.0270
3.0-1.0	0.4633	0.2315	0.1300	0.0781	0.0606	0.0487	0.0339	0.0279	0.0270
2.5-1.5	0.4594	0.2311	0.1298	0.0781	0.0606	0.0486	0.0339	0.0279	0.0270
2.0-2.0	0.4571	0.2306	0.1297	0.0780	0.0606	0.0486	0.0339	0.0279	0.0270
1.5-2.5	0.4533	0.2305	0.1297	0.0780	0.0606	0.0486	0.0339	0.0279	0.0270
1.0-3.0	0.4500	0.2303	0.1294	0.0780	0.0606	0.0486	0.0339	0.0279	0.0270
0.5-3.5	0.4484	0.2301	0.1291	0.0777	0.0606	0.0486	0.0339	0.0279	0.0270

9.2. BRICS

Table S9: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and Atom Pairs torsions fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5460	0.5184	0.4858	0.4384	0.4271	0.4098	0.4085	0.4079	0.4079
3.0-1.0	0.5409	0.5123	0.4810	0.4372	0.4274	0.4099	0.4085	0.4079	0.4079
2.5-1.5	0.5306	0.5030	0.4741	0.4358	0.4286	0.4096	0.4085	0.4079	0.4079
2.0-2.0	0.5216	0.4965	0.4684	0.4337	0.4274	0.4098	0.4084	0.4079	0.4079
1.5-2.5	0.5124	0.4900	0.4632	0.4326	0.4269	0.4098	0.4085	0.4079	0.4079
1.0-3.0	0.5036	0.4863	0.4614	0.4297	0.4259	0.4091	0.4085	0.4079	0.4079
0.5-3.5	0.4966	0.4823	0.4589	0.4278	0.4242	0.4088	0.4085	0.4079	0.4079

Table S10: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and DAIM fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5224	0.5165	0.5122	0.5231	0.5167	0.5134	0.5157	0.5026	0.4812
3.0-1.0	0.5237	0.5188	0.5179	0.5243	0.5179	0.5135	0.5149	0.5017	0.4818
2.5-1.5	0.5266	0.5232	0.5256	0.5263	0.5206	0.5150	0.5130	0.5009	0.4812
2.0-2.0	0.5273	0.5276	0.5271	0.5260	0.5231	0.5164	0.5133	0.5015	0.4814
1.5-2.5	0.5242	0.5267	0.5274	0.5255	0.5201	0.5158	0.5131	0.5034	0.4832
1.0-3.0	0.5232	0.5270	0.5285	0.5224	0.5164	0.5125	0.5080	0.5040	0.4904
0.5-3.5	0.5272	0.5271	0.5260	0.5181	0.5128	0.5100	0.5043	0.5002	0.4894

Table S11: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and ECFP fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5263	0.4894	0.4341	0.4180	0.4163	0.4082	0.4078	0.4078	0.4076
3.0-1.0	0.5212	0.4904	0.4353	0.4183	0.4163	0.4082	0.4078	0.4078	0.4076
2.5-1.5	0.5158	0.4877	0.4371	0.4186	0.4159	0.4082	0.4078	0.4078	0.4076
2.0-2.0	0.5103	0.4868	0.4367	0.4175	0.4145	0.4082	0.4078	0.4078	0.4076
1.5-2.5	0.5058	0.4842	0.4352	0.4169	0.4134	0.4082	0.4078	0.4078	0.4076
1.0-3.0	0.5022	0.4818	0.4345	0.4172	0.4135	0.4082	0.4078	0.4078	0.4076
0.5-3.5	0.4970	0.4770	0.4337	0.4163	0.4126	0.4082	0.4078	0.4078	0.4076

Table S12: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and FCFP fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5271	0.4932	0.4752	0.4661	0.4625	0.4351	0.4345	0.4347	0.4358
3.0-1.0	0.5221	0.5009	0.4832	0.4706	0.4658	0.4365	0.4353	0.4354	0.4358
2.5-1.5	0.5198	0.5053	0.4869	0.4728	0.4687	0.4375	0.4357	0.4358	0.4358
2.0-2.0	0.5181	0.5037	0.4861	0.4718	0.4673	0.4375	0.4359	0.4359	0.4358
1.5-2.5	0.5132	0.4999	0.4847	0.4707	0.4662	0.4380	0.4367	0.4367	0.4358
1.0-3.0	0.5050	0.4980	0.4821	0.4680	0.4643	0.4382	0.4372	0.4371	0.4358
0.5-3.5	0.5002	0.4952	0.4799	0.4658	0.4615	0.4377	0.4362	0.4361	0.4358

Table S13: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and MACCS keys fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5315	0.5392	0.5255	0.4914	0.4769	0.4417	0.4333	0.4143	0.4073
3.0-1.0	0.5343	0.5368	0.5181	0.4922	0.4744	0.4410	0.4322	0.4143	0.4074
2.5-1.5	0.5365	0.5335	0.5191	0.4941	0.4744	0.4425	0.4313	0.4142	0.4074
2.0-2.0	0.5348	0.5285	0.5137	0.4929	0.4716	0.4413	0.4311	0.4145	0.4075
1.5-2.5	0.5344	0.5229	0.5076	0.4894	0.4678	0.4380	0.4300	0.4148	0.4076
1.0-3.0	0.5365	0.5180	0.5017	0.4863	0.4627	0.4329	0.4287	0.4150	0.4076
0.5-3.5	0.5352	0.5168	0.4978	0.4840	0.4610	0.4324	0.4272	0.4152	0.4079

Table S14: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and Topological fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5422	0.5543	0.5523	0.5125	0.4349	0.4187	0.4106	0.4077	0.4078
3.0-1.0	0.5437	0.5497	0.5466	0.5069	0.4351	0.4185	0.4106	0.4077	0.4078
2.5-1.5	0.5435	0.5443	0.5417	0.5026	0.4370	0.4201	0.4105	0.4079	0.4079
2.0-2.0	0.5445	0.5396	0.5308	0.5005	0.4379	0.4209	0.4103	0.4078	0.4079
1.5-2.5	0.5457	0.5339	0.5237	0.4935	0.4390	0.4216	0.4106	0.4079	0.4079
1.0-3.0	0.5444	0.5240	0.5166	0.4893	0.4390	0.4220	0.4100	0.4077	0.4079
0.5-3.5	0.5404	0.5132	0.5047	0.4864	0.4389	0.4225	0.4102	0.4080	0.4079

Table S15: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and Topological torsions fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5134	0.3582	0.2778	0.2547	0.2452	0.2388	0.2192	0.2155	0.2129
3.0-1.0	0.5032	0.3565	0.2777	0.2549	0.2452	0.2389	0.2193	0.2155	0.2129
2.5-1.5	0.4984	0.3549	0.2775	0.2550	0.2453	0.2389	0.2192	0.2155	0.2129
2.0-2.0	0.4947	0.3531	0.2771	0.2549	0.2453	0.2388	0.2192	0.2155	0.2129
1.5-2.5	0.4901	0.3521	0.2766	0.2544	0.2451	0.2386	0.2192	0.2155	0.2129
1.0-3.0	0.4873	0.3520	0.2766	0.2545	0.2450	0.2386	0.2192	0.2155	0.2129
0.5-3.5	0.4829	0.3517	0.2766	0.2545	0.2450	0.2386	0.2192	0.2155	0.2129

9.3. DAIM

Table S16: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and Atom Pairs fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5558	0.5246	0.4847	0.3973	0.3730	0.3346	0.3305	0.3298	0.3291
3.0-1.0	0.5494	0.5221	0.4780	0.3962	0.3726	0.3344	0.3303	0.3298	0.3291
2.5-1.5	0.5398	0.5134	0.4715	0.3946	0.3715	0.3345	0.3303	0.3298	0.3291
2.0-2.0	0.5280	0.5050	0.4678	0.3929	0.3715	0.3342	0.3303	0.3298	0.3291
1.5-2.5	0.5222	0.5003	0.4670	0.3917	0.3707	0.3338	0.3303	0.3298	0.3291
1.0-3.0	0.5159	0.4985	0.4643	0.3902	0.3707	0.3337	0.3303	0.3298	0.3291
0.5-3.5	0.5094	0.4961	0.4612	0.3886	0.3701	0.3337	0.3303	0.3298	0.3291

Table S17: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and DAIM fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5113	0.5139	0.5150	0.5275	0.5343	0.5345	0.5174	0.5165	0.4834
3.0-1.0	0.5143	0.5181	0.5194	0.5281	0.5353	0.5346	0.5188	0.5164	0.4851
2.5-1.5	0.5168	0.5220	0.5250	0.5315	0.5320	0.5327	0.5205	0.5158	0.4871
2.0-2.0	0.5226	0.5295	0.5309	0.5327	0.5312	0.5297	0.5222	0.5160	0.4865
1.5-2.5	0.5291	0.5344	0.5344	0.5307	0.5288	0.5264	0.5212	0.5161	0.4873
1.0-3.0	0.5355	0.5391	0.5340	0.5304	0.5233	0.5204	0.5195	0.5134	0.4869
0.5-3.5	0.5361	0.5374	0.5339	0.5279	0.5173	0.5133	0.5147	0.5046	0.4794

Table S18: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and ECFP fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5377	0.4996	0.3946	0.3487	0.3410	0.3269	0.3260	0.3260	0.3259
3.0-1.0	0.5316	0.4963	0.3935	0.3479	0.3409	0.3268	0.3259	0.3259	0.3259
2.5-1.5	0.5268	0.4919	0.3932	0.3478	0.3407	0.3268	0.3259	0.3259	0.3259
2.0-2.0	0.5200	0.4880	0.3922	0.3478	0.3409	0.3268	0.3259	0.3259	0.3259
1.5-2.5	0.5139	0.4868	0.3919	0.3479	0.3409	0.3269	0.3259	0.3259	0.3259
1.0-3.0	0.5074	0.4855	0.3910	0.3478	0.3409	0.3269	0.3260	0.3259	0.3259
0.5-3.5	0.5002	0.4838	0.3899	0.3473	0.3409	0.3269	0.3260	0.3259	0.3259

Table S19: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and FCFP fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5358	0.5357	0.5169	0.4666	0.4466	0.3829	0.3730	0.3724	0.3654
3.0-1.0	0.5382	0.5310	0.5090	0.4659	0.4471	0.3818	0.3723	0.3718	0.3654
2.5-1.5	0.5342	0.5221	0.4995	0.4654	0.4457	0.3808	0.3714	0.3711	0.3654
2.0-2.0	0.5297	0.5171	0.4987	0.4648	0.4447	0.3806	0.3711	0.3706	0.3654
1.5-2.5	0.5250	0.5113	0.4931	0.4655	0.4445	0.3806	0.3710	0.3704	0.3654
1.0-3.0	0.5199	0.5056	0.4860	0.4641	0.4445	0.3801	0.3707	0.3701	0.3654
0.5-3.5	0.5150	0.5008	0.4840	0.4622	0.4432	0.3799	0.3705	0.3699	0.3654

Table S20: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and MACCS keys fingerprint.

	similarity threshold								
G_C - F_C	0.5385	0.5382	0.5385	0.5129	0.4614	0.3977	0.3712	0.3374	0.3260
3.5-0.5	0.5397	0.5399	0.5359	0.5110	0.4611	0.3988	0.3710	0.3377	0.3261
3.0-1.0	0.5410	0.5388	0.5319	0.5075	0.4603	0.3991	0.3703	0.3376	0.3262
2.5-1.5	0.5384	0.5382	0.5265	0.5024	0.4548	0.3958	0.3692	0.3378	0.3263
2.0-2.0	0.5405	0.5343	0.5211	0.4989	0.4503	0.3910	0.3673	0.3377	0.3264
1.5-2.5	0.5375	0.5309	0.5165	0.4949	0.4456	0.3893	0.3660	0.3378	0.3264
1.0-3.0	0.5329	0.5288	0.5118	0.4905	0.4435	0.3880	0.3655	0.3363	0.3262
0.5-3.5	0.5385	0.5382	0.5385	0.5129	0.4614	0.3977	0.3712	0.3374	0.3260

Table S21: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and Topological fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5299	0.5663	0.5564	0.5231	0.3843	0.3453	0.3354	0.3299	0.3291
3.0-1.0	0.5319	0.5636	0.5533	0.5227	0.3831	0.3454	0.3356	0.3300	0.3291
2.5-1.5	0.5381	0.5521	0.5499	0.5199	0.3821	0.3453	0.3357	0.3301	0.3291
2.0-2.0	0.5407	0.5529	0.5439	0.5073	0.3816	0.3443	0.3357	0.3301	0.3291
1.5-2.5	0.5463	0.5431	0.5382	0.5147	0.3809	0.3432	0.3351	0.3301	0.3291
1.0-3.0	0.5502	0.5345	0.5306	0.5105	0.3801	0.3431	0.3352	0.3302	0.3291
0.5-3.5	0.5474	0.5250	0.5197	0.5063	0.3796	0.3432	0.3353	0.3302	0.3292

Table S22: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class with RECAP fragmentation rules and Topological torsions fingerprint.

	similarity threshold								
G_C - F_C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.5-0.5	0.5080	0.3872	0.3113	0.2846	0.2751	0.2694	0.2448	0.2448	0.2442
3.0-1.0	0.5015	0.3858	0.3113	0.2845	0.2755	0.2697	0.2448	0.2448	0.2442
2.5-1.5	0.4942	0.3815	0.3111	0.2845	0.2756	0.2698	0.2448	0.2448	0.2442
2.0-2.0	0.4900	0.3793	0.3111	0.2847	0.2758	0.2699	0.2448	0.2448	0.2442
1.5-2.5	0.4857	0.3780	0.3110	0.2846	0.2756	0.2698	0.2448	0.2448	0.2442
1.0-3.0	0.4841	0.3780	0.3104	0.2846	0.2756	0.2698	0.2448	0.2448	0.2442
0.5-3.5	0.4821	0.3778	0.3101	0.2844	0.2754	0.2696	0.2448	0.2448	0.2442

9.4. Globals

Table S23: Average median AUC values derived from AUC-distributions obtained from 30 retrieval runs per activity class without our algorithmic graph reduction.

fingerprint	average median AUC
apairs	0.6298
daim	0.5272
ecfp	0.5838
fcfp	0.5651
maccs	0.5600
topo	0.5440
torsion	0.6224

9.5. Length of scoring lists

Table S24: Average median length of scoring lists obtained from 30 retrieval runs per activity class with Atom Pairs fingerprint.

	similarity threshold								
rule	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
RECAP	13411	8212	4084	1696	1088	664	568	551	551
BRICS	14627	13931	12113	9947	9616	8989	8941	8930	8929
DAIM	14530	13230	11424	7401	6979	5602	5521	5505	5498

Table S25: Average median length of scoring lists obtained from 30 retrieval runs per activity class with DAIM fingerprint.

	similarity threshold								
rule	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
RECAP	15027	15027	15027	15022	14970	14681	13810	11963	8032
BRICS	15027	15027	15026	15019	14990	14897	14674	14145	13315
DAIM	15027	15027	15027	15025	15010	14949	14704	13964	12200

Table S26: Average median length of scoring lists obtained from 30 retrieval runs per activity class with ECFP fingerprint.

	similarity threshold								
rule	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
RECAP	13429	5942	2027	1010	716	560	555	551	546
BRICS	14750	13845	10060	9171	9038	8888	8887	8887	8887
DAIM	14680	13163	7809	6352	5916	5460	5445	5445	5444

Table S27: Average median length of scoring lists obtained from 30 retrieval runs per activity class with FCFP fingerprint.

	similarity threshold								
rule	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
RECAP	14963	13759	8880	4643	2910	1278	782	683	604
BRICS	14973	14582	13029	12040	11589	10107	9997	9995	9877
DAIM	15010	14765	13564	11666	10423	7311	7097	7075	6860

Table S28: Average median length of scoring lists obtained from 30 retrieval runs per activity class with MACCS Keys fingerprint.

	similarity threshold								
rule	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
RECAP	15026	14720	12174	7807	4422	2338	1464	892	629
BRICS	15026	14917	14346	13574	12082	9954	9567	9057	8917
DAIM	15027	14947	14099	12393	9765	7588	6892	5666	5464

Table S29: Average median length of scoring lists obtained from 30 retrieval runs per activity class with Topological fingerprint (corresponds to **Error! Reference source not found.B** (topo)).

	similarity threshold								
rule	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
RECAP	15027	15025	14959	11313	2670	1063	774	627	561
BRICS	15027	15026	14979	13652	10226	9514	9018	8960	8930
DAIM	15027	15024	14976	13358	7106	5904	5624	5571	5508

Table S30: Average median length of scoring lists obtained from 30 retrieval runs per activity class with Torsions fingerprint (corresponds to **Error! Reference source not found.B** (torsions)).

	similarity threshold								
rule	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
RECAP	7995	2808	1175	651	487	333	243	219	216
BRICS	11195	6428	4383	3557	3399	3226	2428	2399	2294
DAIM	11893	7158	5228	4422	4149	4004	3185	3141	3116

9.6. Statistical significance of RedFrag runs at OOPS and run at best single parameter set

Table S31: Statistical significance of best performances for each activity class calculated for AUC distributions resulting from RedFrag runs at OOPS and the best single parameter set (ks-test: Kolmogorov-Smirnov test).

	MACCS|0.1|0.5-3.5		RedFrag at OOPS			
target	median	mean	median	mean	p-value (t-test)	p-value (ks-test)
S1P1 receptor	0.570	0.567	0.682	0.669	0.0000	0.0000
PKA	0.650	0.656	0.657	0.635	0.3930	0.5372
SF1 inhibitors	0.501	0.496	0.596	0.583	0.0000	0.0000
Rho-Kinase2	0.607	0.600	0.607	0.600	1.0000	1.0000
HIV RT-RNase	0.513	0.495	0.632	0.586	0.0001	0.0000
EPH receptor A4	0.542	0.526	0.558	0.550	0.1446	0.1088
SF1 agonists	0.529	0.532	0.665	0.641	0.0000	0.0000
HSP 90	0.597	0.591	0.626	0.600	0.7750	0.5372
ER-α-coactivator binding inhibitor	0.499	0.499	0.557	0.531	0.0574	0.0259
ER-β-coactivator binding inhibitor	0.519	0.523	0.583	0.539	0.4484	0.1088
ER-α-coactivator binding potentiator	0.605	0.596	0.628	0.600	0.8363	0.5372
FAK	0.581	0.564	0.628	0.538	0.3938	0.0046
Cathepsin G	0.661	0.658	0.661	0.660	0.9094	0.7600
FXIa	0.708	0.695	0.780	0.721	0.3989	0.0259
FXIIa	0.748	0.705	0.762	0.713	0.7923	0.7600
D1 receptor	0.583	0.553	0.630	0.570	0.5972	0.0046
M1 receptor	0.556	0.533	0.571	0.531	0.9541	0.5372

9.7. RedFrag performance at OOPS

Table S32: Best performing sets of parameters for each activity class from the MUV data sets.

target	median	mean	fingerprint	fragmentation	similarity threshold	G_C - F_C
S1P1 receptor	0.682	0.669	topo	RECAP	0.2	3.5-0.5
PKA	0.657	0.635	maccs	DAIM	0.1	1.5-2.5
SF1 inhibitors	0.596	0.583	topo	RECAP	0.3	3.5-0.5
Rho-Kinase2	0.607	0.600	maccs	RECAP	0.1	0.5-3.5
HIV RT-RNase	0.632	0.586	topo	DAIM	0.2	3.5-0.5
EPH receptor A4	0.558	0.550	apairs	DAIM	0.1	3.0-1.0
SF1 agonists	0.665	0.641	daim	BRICS	0.1	3.5-0.5
HSP 90	0.626	0.600	maccs	RECAP	0.2	3.5-0.5
ER-α-coactivator binding inhibitor	0.557	0.531	daim	RECAP	0.3	0.5-3.5
ER-β-coactivator binding inhibitor	0.583	0.539	maccs	RECAP	0.1	3.5-0.5
ER-α-coactivator binding potentiator	0.628	0.600	maccs	RECAP	0.2	2.0-2.0
FAK	0.628	0.538	topo	BRICS	0.3	3.0-1.0
Cathepsin G	0.661	0.660	maccs	RECAP	0.2	0.5-3.5
FXIa	0.780	0.721	topo	DAIM	0.2	3.5-0.5
FXIIa	0.762	0.713	maccs	RECAP	0.1	1.0-3.0
D1 receptor	0.630	0.570	torsions	BRICS	0.1	3.5-0.5
M1 receptor	0.571	0.531	maccs	RECAP	0.3	1.0-3.0

9.8. Statistical significance of performance restricted to different fragmentation rules

Table S33: Statistical significance of best performances for each activity class calculated for AUC distributions resulting from RECAP and DAIM fragmentation rules (ks-test: Kolmogorov-Smirnov test).

	RECAP		DAIM			
target	median	mean	median	mean	p-value (t-test)	p-value (ks-test)
S1P1 receptor	0.682	0.669	0.532	0.525	0.0000	0.0000
PKA	0.654	0.650	0.657	0.635	0.5783	0.7600
SF1 inhibitors	0.596	0.583	0.538	0.529	0.0000	0.0000
Rho-Kinase2	0.607	0.600	0.599	0.576	0.2829	0.2003
HIV RT-RNase	0.536	0.513	0.632	0.586	0.0005	0.0000
EPH receptor A4	0.542	0.543	0.558	0.550	0.6560	0.7600
SF1 agonists	0.558	0.540	0.584	0.576	0.0083	0.0259
HSP 90	0.626	0.600	0.534	0.514	0.0019	0.0001
ER-α-coactivator binding inhibitor	0.557	0.531	0.548	0.527	0.7916	0.5372
ER-β-coactivator binding inhibitor	0.583	0.539	0.550	0.539	0.9964	0.0259
ER-α-coactivator binding potentiator	0.628	0.600	0.598	0.608	0.6147	0.3420
FAK	0.602	0.575	0.552	0.531	0.0094	0.0006
Cathepsin G	0.661	0.660	0.622	0.576	0.0003	0.0002
FXIa	0.712	0.696	0.780	0.721	0.4230	0.0550
FXIIa	0.762	0.713	0.723	0.669	0.1692	0.0046
D1 receptor	0.598	0.541	0.626	0.588	0.0951	0.2003
M1 receptor	0.571	0.531	0.570	0.539	0.7873	0.9360

Table S34: Statistical significance of best performances for each activity class calculated for AUC distributions resulting from DAIM and BRICS fragmentation rules (ks-test: Kolmogorov-Smirnov test).

	DAIM		BRICS			
target	median	mean	median	mean	p-value (t-test)	p-value (ks-test)
S1P1 receptor	0.532	0.525	0.521	0.514	0.4539	0.1088
PKA	0.657	0.635	0.623	0.599	0.1789	0.5372
SF1 inhibitors	0.538	0.529	0.545	0.532	0.8215	0.9360
Rho-Kinase2	0.599	0.576	0.570	0.563	0.4368	0.3420
HIV RT-RNase	0.632	0.586	0.544	0.528	0.0430	0.0550
EPH receptor A4	0.558	0.550	0.552	0.530	0.2888	0.5372
SF1 agonists	0.584	0.576	0.665	0.641	0.0000	0.0000
HSP 90	0.534	0.514	0.576	0.557	0.1514	0.0550
ER-α-coactivator binding inhibitor	0.548	0.527	0.521	0.513	0.4664	0.3420
ER-β-coactivator binding inhibitor	0.550	0.539	0.536	0.525	0.2182	0.0550
ER-α-coactivator binding potentiator	0.598	0.608	0.593	0.582	0.1266	0.3420
FAK	0.552	0.531	0.628	0.538	0.8358	0.0006
Cathepsin G	0.622	0.576	0.636	0.629	0.0115	0.0259
FXIa	0.780	0.721	0.745	0.680	0.2918	0.0046
FXIIa	0.723	0.669	0.676	0.628	0.2616	0.1088
D1 receptor	0.626	0.588	0.630	0.570	0.5588	0.5372
M1 receptor	0.570	0.539	0.540	0.534	0.8301	0.0259

Table S35: Statistical significance of best performances for each activity class calculated for AUC distributions resulting from RECAP and BRICS fragmentation rules (ks-test: Kolmogorov-Smirnov test).

	RECAP		BRICS			
target	median	mean	median	mean	p-value (t-test)	p-value (ks-test)
S1P1 receptor	0.682	0.669	0.521	0.514	0.0000	0.0000
PKA	0.654	0.650	0.623	0.599	0.0682	0.3420
SF1 inhibitors	0.596	0.583	0.545	0.532	0.0001	0.0000
Rho-Kinase2	0.607	0.600	0.570	0.563	0.0839	0.1088
HIV RT-RNase	0.536	0.513	0.544	0.528	0.5681	0.0020
EPH receptor A4	0.542	0.543	0.552	0.530	0.4853	0.5372
SF1 agonists	0.558	0.540	0.665	0.641	0.0000	0.0000
HSP 90	0.626	0.600	0.576	0.557	0.2023	0.0259
ER-α-coactivator binding inhibitor	0.557	0.531	0.521	0.513	0.3096	0.2003
ER-β-coactivator binding inhibitor	0.583	0.539	0.536	0.525	0.4084	0.0006
ER-α-coactivator binding potentiator	0.628	0.600	0.593	0.582	0.3477	0.2003
FAK	0.602	0.575	0.628	0.538	0.2134	0.0046
Cathepsin G	0.661	0.660	0.636	0.629	0.0415	0.0550
FXIa	0.712	0.696	0.745	0.680	0.6380	0.3420
FXIIa	0.762	0.713	0.676	0.628	0.0190	0.0002
D1 receptor	0.598	0.541	0.630	0.570	0.4321	0.0550
M1 receptor	0.571	0.531	0.540	0.534	0.9022	0.0259

9.9. Best performances of fingerprints without graph reduction

Table S36: Best performing fingerprints without algorithmic graph reduction for each activity class from the MUV data sets.

target	median	mean	fingerprint
S1P1 receptor	0.594	0.574	apairs
PKA	0.710	0.693	apairs
SF1 inhibitors	0.630	0.623	apairs
Rho-Kinase2	0.607	0.598	torsions
HIV RT-RNase	0.615	0.607	torsions
EPH receptor A4	0.643	0.622	torsions
SF1 agonists	0.561	0.554	apairs
HSP 90	0.735	0.705	torsions
ER-α-coactivator binding inhibitor	0.558	0.548	torsions
ER-β-coactivator binding inhibitor	0.560	0.550	apairs
ER-α-coactivator binding potentiator	0.742	0.728	apairs
FAK	0.647	0.643	torsions
Cathepsin G	0.686	0.681	maccs
FXIa	0.790	0.776	apairs
FXIIa	0.757	0.745	apairs
D1 receptor	0.574	0.578	topo
M1 receptor	0.589	0.558	ecfp

9.10. Best performances of RedFrag restricted to a single fragmentation rule

Table S37: Best performing sets of parameters for each activity class from the MUV data sets corresponding to the RECAP fragmentation rules.

target	median	mean	fingerprint	similarity threshold	G_C - F_C
S1P1 receptor	0.682	0.669	topo	0.2	3.5-0.5
PKA	0.654	0.650	maccs	0.2	0.5-3.5
SF1 inhibitors	0.596	0.583	topo	0.3	3.5-0.5
Rho-Kinase2	0.607	0.600	maccs	0.1	0.5-3.5
HIV RT-RNase	0.536	0.513	topo	0.2	2.0-2.0
EPH receptor A4	0.542	0.543	maccs	0.2	0.5-3.5
SF1 agonists	0.558	0.540	daim	0.6	0.5-3.5
HSP 90	0.626	0.600	maccs	0.2	3.5-0.5
ER-α-coactivator binding inhibitor	0.557	0.531	daim	0.3	0.5-3.5
ER-β-coactivator binding inhibitor	0.583	0.539	maccs	0.1	3.5-0.5
ER-α-coactivator binding potentiator	0.628	0.600	maccs	0.2	2.0-2.0
FAK	0.602	0.575	topo	0.3	2.0-2.0
Cathepsin G	0.661	0.660	maccs	0.2	0.5-3.5
FXIa	0.712	0.696	maccs	0.1	1.0-3.0
FXIIa	0.762	0.713	maccs	0.1	1.0-3.0
D1 receptor	0.598	0.541	daim	0.4	2.0-2.0
M1 receptor	0.571	0.531	maccs	0.3	1.0-3.0

Table S38: Best performing sets of parameters for each activity class from the MUV data sets corresponding to the DAIM fragmentation rules.

target	median	mean	fingerprint	similarity threshold	G_C - F_C
S1P1 receptor	0.532	0.525	daim	0.7	0.5-3.5
PKA	0.657	0.635	maccs	0.1	1.5-2.5
SF1 inhibitors	0.538	0.529	topo	0.3	3.5-0.5
Rho-Kinase2	0.599	0.576	daim	0.3	0.5-3.5
HIV RT-RNase	0.632	0.586	topo	0.2	3.5-0.5
EPH receptor A4	0.558	0.550	apairs	0.1	3.0-1.0
SF1 agonists	0.584	0.576	daim	0.1	3.0-1.0
HSP 90	0.534	0.514	daim	0.5	0.5-3.5
ER-α-coactivator binding inhibitor	0.548	0.527	topo	0.3	3.5-0.5
ER-β-coactivator binding inhibitor	0.550	0.539	topo	0.3	3.5-0.5
ER-α-coactivator binding potentiator	0.598	0.608	apairs	0.1	3.5-0.5
FAK	0.552	0.531	topo	0.2	3.5-0.5
Cathepsin G	0.622	0.576	maccs	0.4	2.5-1.5
FXIa	0.780	0.721	topo	0.2	3.5-0.5
FXIIa	0.723	0.669	topo	0.2	3.5-0.5
D1 receptor	0.626	0.588	maccs	0.3	3.0-1.0
M1 receptor	0.570	0.539	torsions	0.1	3.5-0.5

Table S39: Best performing sets of parameters for each activity class from the MUV data sets corresponding to the BRICS fragmentation rules.

target	median	mean	fingerprint	similarity threshold	G_C - F_C
S1P1 receptor	0.521	0.514	topo	0.3	3.5-0.5
PKA	0.623	0.599	maccs	0.1	2.5-1.5
SF1 inhibitors	0.545	0.532	topo	0.2	2.5-1.5
Rho-Kinase2	0.570	0.563	fcfp	0.1	3.5-0.5
HIV RT-RNase	0.544	0.528	topo	0.3	3.5-0.5
EPH receptor A4	0.552	0.530	daim	0.2	0.5-3.5
SF1 agonists	0.665	0.641	daim	0.1	3.5-0.5
HSP 90	0.576	0.557	apairs	0.1	3.5-0.5
ER-α-coactivator binding inhibitor	0.521	0.513	topo	0.2	3.5-0.5
ER-β-coactivator binding inhibitor	0.536	0.525	daim	0.4	3.5-0.5
ER-α-coactivator binding potentiator	0.593	0.582	fcfp	0.2	3.5-0.5
FAK	0.628	0.538	topo	0.3	3.0-1.0
Cathepsin G	0.636	0.629	maccs	0.1	0.5-3.5
FXIa	0.745	0.680	apairs	0.2	3.5-0.5
FXIIa	0.676	0.628	apairs	0.1	3.5-0.5
D1 receptor	0.630	0.570	torsions	0.1	3.5-0.5
M1 receptor	0.540	0.534	daim	0.5	3.5-0.5

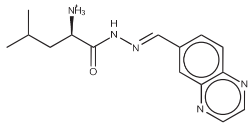
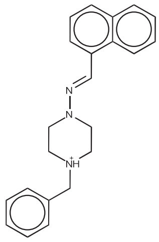
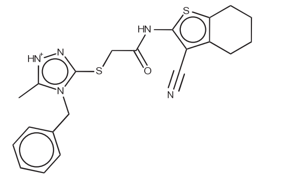
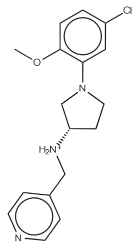
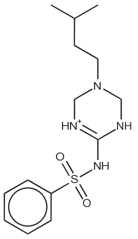
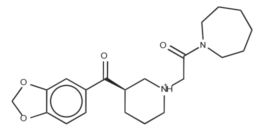
9.11. Statistical significance of RedFrag performance at OOPS and best Globals

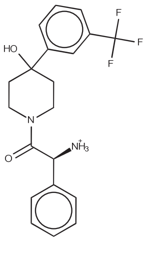
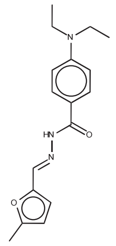
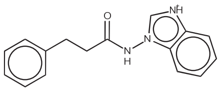
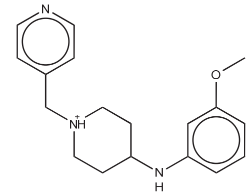
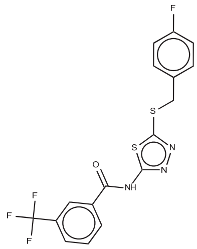
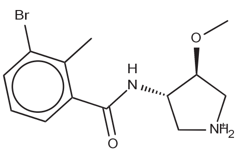
Table S40: Statistical significance of best performances for each activity class calculated for AUC distributions resulting from RedFrag runs and calculations without algorithmic graph reduction (Globals) (ks-test: Kolmogorov-Smirnov test).

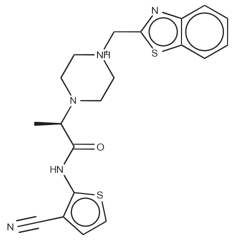
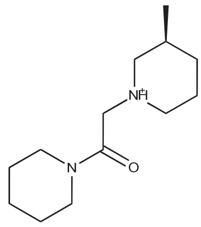
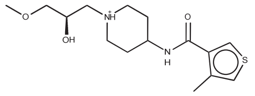
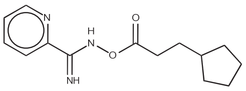
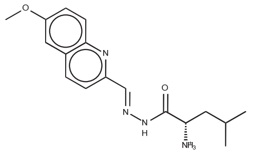
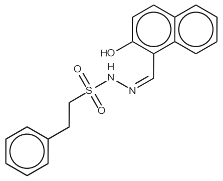
	RedFrag		Globals			
target	median	mean	median	mean	p-value (t-test)	p-value (ks-test)
S1P1 receptor	0.682	0.669	0.594	0.574	0.0000	0.0000
PKA	0.657	0.635	0.710	0.693	0.0144	0.0259
SF1 inhibitors	0.596	0.583	0.630	0.623	0.0047	0.0006
Rho-Kinase2	0.607	0.600	0.607	0.598	0.9276	0.7600
HIV RT-RNase	0.632	0.586	0.615	0.607	0.2920	0.2003
EPH receptor A4	0.558	0.550	0.643	0.622	0.0002	0.0017
SF1 agonists	0.665	0.641	0.561	0.554	0.0000	0.0000
HSP 90	0.626	0.600	0.735	0.705	0.0006	0.0006
ER-α-coactivator binding inhibitor	0.557	0.531	0.558	0.548	0.3041	0.5372
ER-β-coactivator binding inhibitor	0.583	0.539	0.560	0.550	0.5578	0.5372
ER-α-coactivator binding potentiator	0.628	0.600	0.742	0.728	0.0000	0.0000
FAK	0.628	0.538	0.647	0.643	0.0008	0.0113
Cathepsin G	0.661	0.660	0.686	0.681	0.2463	0.3420
FXIa	0.780	0.721	0.790	0.776	0.0482	0.5372
FXIIa	0.762	0.713	0.757	0.745	0.2665	0.1088
D1 receptor	0.630	0.570	0.574	0.578	0.7629	0.0006
M1 receptor	0.571	0.531	0.589	0.558	0.2463	0.3420

10. Virtual hits and query compounds

Table S41: Virtual hits with 2D depiction of the molecular structure, the CXSMILES string as built by ChemAxon’s MarvinSketch, molecule title used throughout the main document.

Structure	CXSMILES	Molecule Title	Original Title
	CC(C)C[C@@H]([NH3+])C(=O)N\N=C\c1ccc2nccnc2c1	1	K00JG014
	C([NH+]1CCN(CC1)\N=C\c1cccc2ccccc12)c1ccccc1	2	K00JG001
	Cc1[nH+]nc(SCC(=O)Nc2sc3CCCCc3c2C#N)n1Cc1ccccc1	3	K00JG013
	COc1ccc(Cl)cc1N1CC[C@@H](C1)[NH2+]Cc1ccncc1	4	K00JG004
	CC(C)CCN1CNC(NS(=O)(=O)c2ccccc2)=[NH+]C1	5	K00JG018
	O=C(C[NH+]1CCC[C@H](C1)C(=O)c1ccc2OCOc2c1)N1CCCCCC1	6	K00JG003

	[NH3+][C@H](C(=O)N1CCC(O)(CC1)c1cccc(c1)C(F)(F)F)c1ccccc1	7	K00JG002
	CCN(CC)c1ccc(cc1)C(=O)N\N=C\c1ccc(C)o1	8	K00JG004
	O=C(CCc1ccccc1)Nn1c[nH+]c2ccccc12	9	K00JG009
	COc1cccc(NC2CC[NH+](Cc3ccncc3)CC2)c1	10	K00JG007
	Fc1ccc(CSc2nnc(NC(=O)c3cccc(c3)C(F)(F)F)s2)cc1	11	K00JG008
	CO[C@H]1C[NH2+]C[C@@H]1NC(=O)c1cccc(Br)c1C	12	K00JG009

	C[C@@H](N1CC[NH+](Cc2nc3ccccc3s2)CC1)C(=O)Nc1sccc1C#N	13	K00JG010
	C[C@H]1CCC[NH+](CC(=O)N2CCCCC2)C1	14	K00JG011
	COC[C@H](O)C[NH+]1CCC(CC1)NC(=O)c1cscc1C	15	K00JG012
	N=C(NOC(=O)CCC1CCCC1)c1ccccn1	16	K00JG015
	COc1ccc2nc(\C=N\NC(=O)[C@@H]([NH3+])CC(C)C)ccc2c1	17	K00JG016
	Oc1ccc2ccccc2c1\C=N/NS(=O)(=O)CCc1ccccc1	18	K00JG017

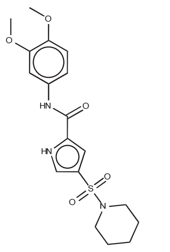
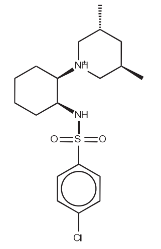
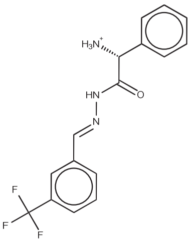
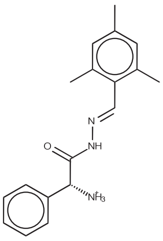
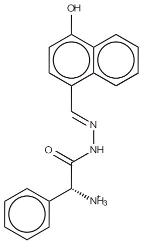
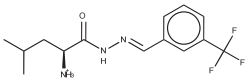
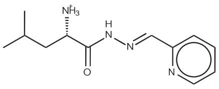
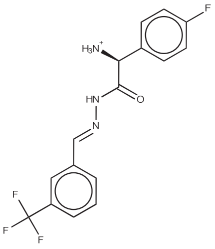
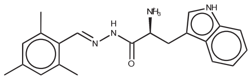
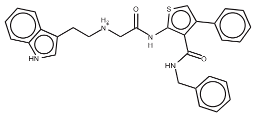
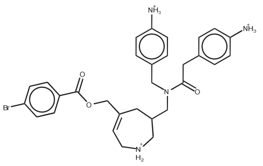
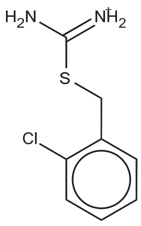
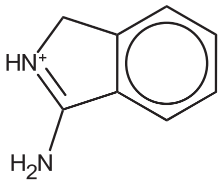
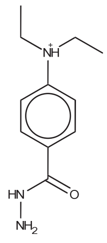
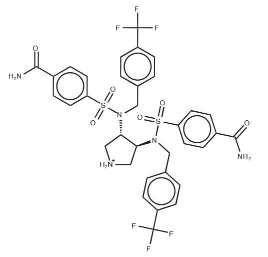
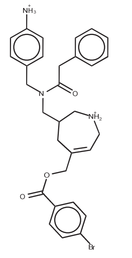
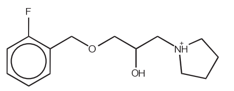
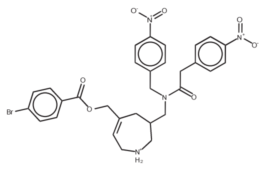
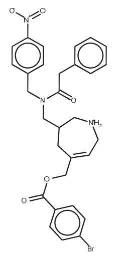
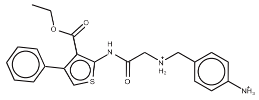
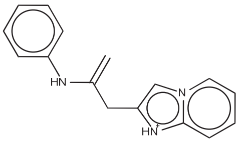
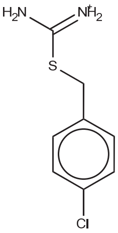
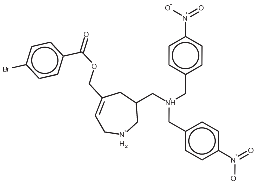
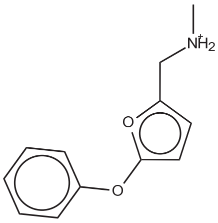
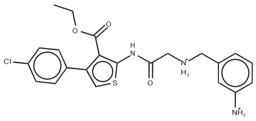
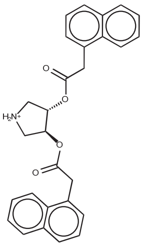
	COc1ccc(NC(=O)c2cc(c[nH]2)S(=O)(=O)N2CCCCC2)cc1OC	19	K00JG019
	C[C@@H]1C[C@@H](C)C[NH+](C1)[C@@H]1CCCC[C@@H]1NS(=O)(=O)c1ccc(Cl)cc1	20	K00JG020

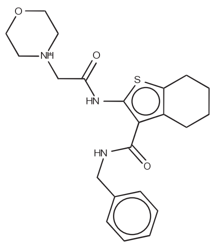
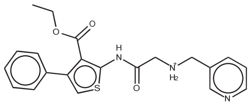
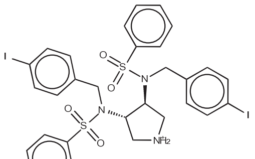
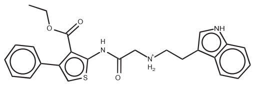
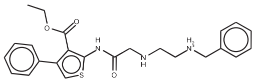
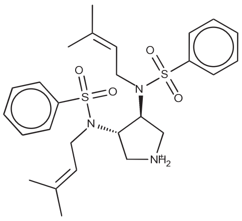
Table S42: Query compounds with 2D depiction of the molecular structure, the CXSMILES string as built by ChemAxon's MarvinSketch, molecule title used throughout the main document.

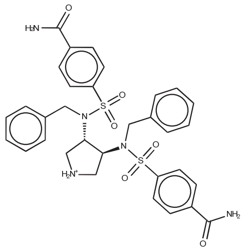
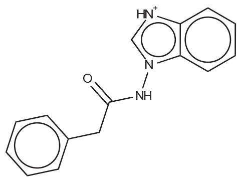
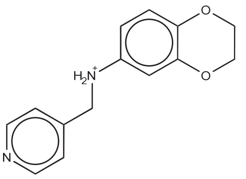
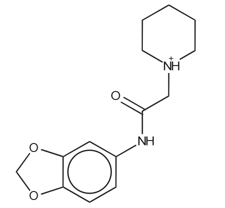
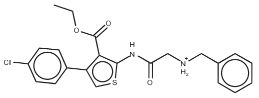
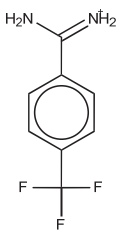
Structure	CXSMILES	Molecule Title	Original Title
	[NH3+][C@@H](C(=O)N\N=C\c1cccc(c1)C(F)(F)F)c1ccccc1	Q1	R_H3_A1
	Cc1cc(C)c(\C=N\NC(=O)[C@H]([NH3+])c2ccccc2)c(C)c1	Q2	R_H3_A4
	[NH3+][C@@H](C(=O)N\N=C\c1ccc(O)c2ccccc12)c1ccccc1	Q3	R_H3_A5
	CC(C)C[C@H]([NH3+])C(=O)N\N=C\c1cccc(c1)C(F)(F)F	Q4	S_H1_A1
	CC(C)C[C@H]([NH3+])C(=O)N\N=C\c1ccccn1	Q5	S_H1_A3
	[NH3+][C@H](C(=O)N\N=C\c1cccc(c1)C(F)(F)F)c1ccc(F)cc1	Q6	S_H2_A1

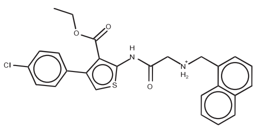
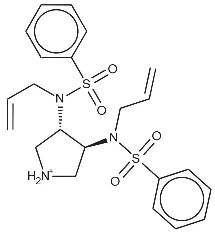
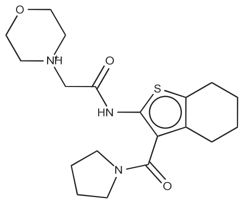
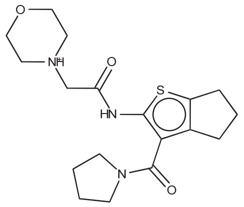
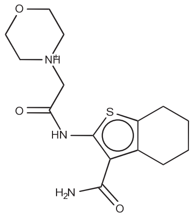
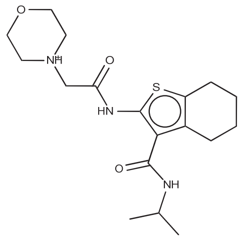
	Cc1cc(C)c(\C=N\NC(=O)[C@@H]([NH3+])Cc2c[nH]c3ccccc23)c(C)c1	Q7	S_H4_A4
	O=C(C[NH2+]CCc1c[nH]c2ccccc12)Nc1scc(c1C(=O)NCc1ccccc1)-c1ccccc1	Q8	RB91
	[NH3+]c1ccc(CN(CC2C[NH2+]CC=C(COC(=O)c3ccc(Br)cc3)C2)C(=O)Cc2ccc([NH3+])cc2)cc1	Q9	NC239
	NC(=[NH2+])SCc1ccccc1Cl	Q10	F306
	NC1=[NH+]Cc2ccccc12	Q11	F005
	CC[NH+](CC)c1ccc(cc1)C(=O)NN	Q12	F109

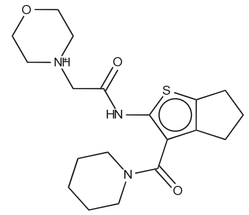
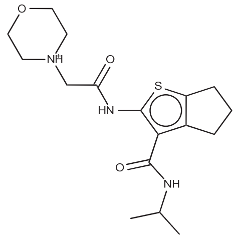
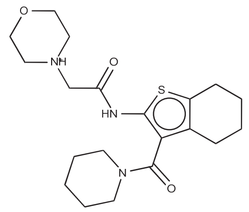
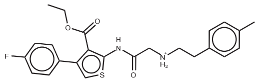
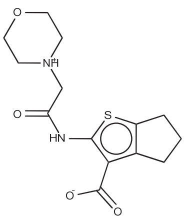
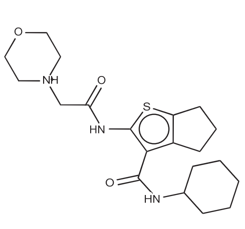
	NC(=O)c1ccc(cc1)S(=O)(=O)N(Cc1ccc(cc1)C(F)(F)F)[C@H]1C[NH2+]C[C@@H]1N(Cc1ccc(cc1)C(F)(F)F)S(=O)(=O)c1ccc(cc1)C(N)=O	Q13	AB111
	[NH3+]c1ccc(CN(CC2C[NH2+]CC=C(COC(=O)c3ccc(Br)cc3)C2)C(=O)Cc2ccccc2)cc1	Q14	ST231
	OC(COCc1ccccc1F)C[NH+]1CCCC1	Q15	F284
	[O-][N+](=O)c1ccc(CN(CC2C[NH2+]CC=C(COC(=O)c3ccc(Br)cc3)C2)C(=O)Cc2ccc(cc2)[N+]([O-])=O)cc1	Q16	NC230
	[O-][N+](=O)c1ccc(CN(CC2C[NH2+]CC=C(COC(=O)c3ccc(Br)cc3)C2)C(=O)Cc2ccccc2)cc1	Q17	NC231
	CCOC(=O)c1c(NC(=O)C[NH2+]Cc2ccc([NH3+])cc2)scc1-c1ccccc1	Q18	RB51

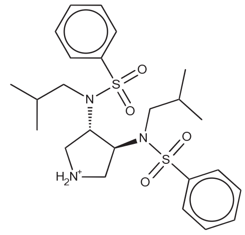
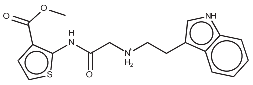
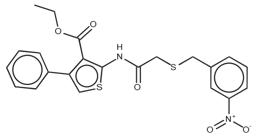
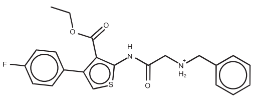
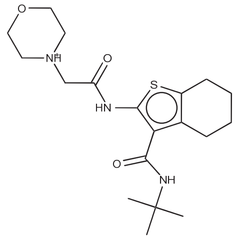
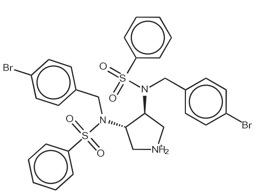
	C=C(Cc1cn2ccccc2[nH+]1)Nc1ccccc1	Q19	F255
	NC(=[NH2+])SCc1ccc(Cl)cc1	Q20	F290
	[O-][N+](=O)c1ccc(C[NH+](CC2C[NH2+]CC=C(COC(=O)c3ccc(Br)cc3)C2)Cc2ccc(cc2)[N+]([O-])=O)cc1	Q21	NC229
	C[NH2+]Cc1ccc(Oc2ccccc2)o1	Q22	F063
	CCOC(=O)c1c(NC(=O)C[NH2+]Cc2cccc([NH3+])c2)scc1-c1ccc(Cl)cc1	Q23	RB57
	O=C(Cc1cccc2ccccc12)O[C@H]1C[NH2+]C[C@@H]1OC(=O)Cc1cccc2ccccc12	Q24	ST47

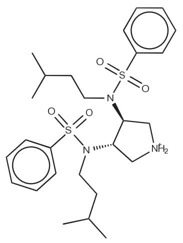
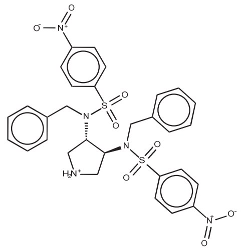
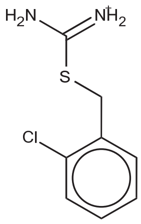
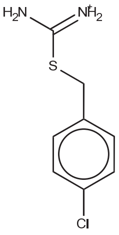
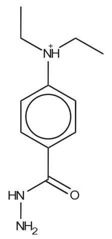
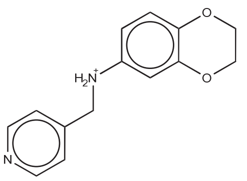
	O=C(C[NH+]1CCOCC1)Nc1sc2CCCCc2c1C(=O)NCc1ccccc1	Q25	SH40
	CCOC(=O)c1c(NC(=O)C[NH2+]Cc2cccnc2)scc1-c1ccccc1	Q26	RB50
	Ic1ccc(CN([C@H]2C[NH2+]C[C@@H]2N(Cc2ccc(I)cc2)S(=O)(=O)c2ccccc2)S(=O)(=O)c2ccccc2)cc1	Q27	AB88
	CCOC(=O)c1c(NC(=O)C[NH2+]CCc2c[nH]c3ccccc23)scc1-c1ccccc1	Q28	RB49
	CCOC(=O)c1c(NC(=O)CNCC[NH2+]Cc2ccccc2)scc1-c1ccccc1	Q29	RB48
	CC(C)=CCN([C@H]1C[NH2+]C[C@@H]1N(CC=C(C)C)S(=O)(=O)c1ccccc1)S(=O)(=O)c1ccccc1	Q30	AB102

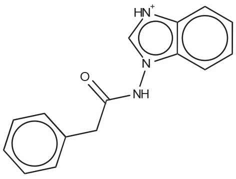
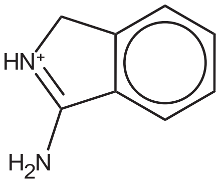
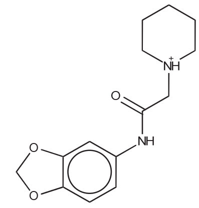
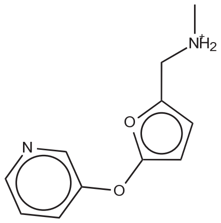
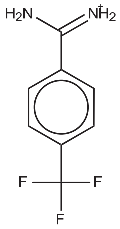
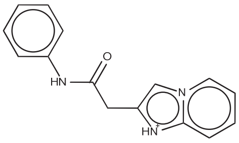
	NC(=O)c1ccc(cc1)S(=O)(=O)N(Cc1ccccc1)[C@H]1C[NH2+]C[C@@H]1N(Cc1ccccc1)S(=O)(=O)c1ccc(cc1)C(N)=O	Q31	AB99
	O=C(Cc1ccccc1)Nn1c[nH+]c2ccccc12	Q32	F148
	C([NH2+]c1ccc2OCCOc2c1)c1ccncc1	Q33	F291
	O=C(C[NH+]1CCCCC1)Nc1ccc2OCOc2c1	Q34	F041
	CCOC(=O)c1c(NC(=O)C[NH2+]Cc2ccccc2)scc1-c1ccc(Cl)cc1	Q35	RB30
	NC(=[NH2+])c1ccc(cc1)C(F)(F)F	Q36	F216

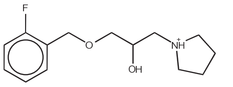
	CCOC(=O)c1c(NC(=O)C[NH2+]Cc2cccc3ccccc23)scc1-c1ccc(Cl)cc1	Q37	RB34
	C=CCN([C@H]1C[NH2+]C[C@@H]1N(CC=C)S(=O)(=O)c1ccccc1)S(=O)(=O)c1ccccc1	Q38	AB100
	O=C(C[NH+]1CCOCC1)Nc1sc2CCCCc2c1C(=O)N1CCCC1	Q39	SH33
	O=C(C[NH+]1CCOCC1)Nc1sc2CCCc2c1C(=O)N1CCCC1	Q40	AM16
	NC(=O)c1c(NC(=O)C[NH+]2CCOCC2)sc2CCCCc12	Q41	SH42
	CC(C)NC(=O)c1c(NC(=O)C[NH+]2CCOCC2)sc2CCCCc12	Q42	SH43

	O=C(C[NH+]1CCOCC1)Nc1sc2CCCc2c1C(=O)N1CCCCC1	Q43	AM7
	CC(C)NC(=O)c1c(NC(=O)C[NH+]2CCOCC2)sc2CCCc12	Q44	AM18
	O=C(C[NH+]1CCOCC1)Nc1sc2CCCCc2c1C(=O)N1CCCCC1	Q45	SH36
	CCOC(=O)c1c(NC(=O)C[NH2+]CCc2ccc(C)cc2)scc1-c1ccc(F)cc1	Q46	RB33
	[O-]C(=O)c1c(NC(=O)C[NH+]2CCOCC2)sc2CCCc12	Q47	AM6
	O=C(C[NH+]1CCOCC1)Nc1sc2CCCc2c1C(=O)NC1CCCCC1	Q48	AM8

	CC(C)CN([C@H]1C[NH2+]C[C@@H]1N(CC(C)C)S(=O)(=O)c1ccccc1)S(=O)(=O)c1ccccc1	Q49	AB115
	COC(=O)c1ccsc1NC(=O)C[NH2+]CCc1c[nH]c2ccccc12	Q50	RB73
	CCOC(=O)c1c(NC(=O)CSCc2cccc(c2)[N+]([O-])=O)scc1-c1ccccc1	Q51	RB66
	CCOC(=O)c1c(NC(=O)C[NH2+]Cc2ccccc2)scc1-c1ccc(F)cc1	Q52	RB31
	CC(C)(C)NC(=O)c1c(NC(=O)C[NH+]2CCOCC2)sc2CCCCc12	Q53	SH41
	Brc1ccc(CN([C@H]2C[NH2+]C[C@@H]2N(Cc2ccc(Br)cc2)S(=O)(=O)c2ccccc2)S(=O)(=O)c2ccccc2)cc1	Q54	AB86

	CC(C)CCN([C@H]1C[NH2+]C[C@@H]1N(CCC(C)C)S(=O)(=O)c1ccccc1)S(=O)(=O)c1ccccc1	Q55	AB116
	[O-][N+](=O)c1ccc(cc1)S(=O)(=O)N(Cc1ccccc1)[C@H]1C[NH2+]C[C@@H]1N(Cc1ccccc1)S(=O)(=O)c1ccc(cc1)[N+]([O-])=O	Q56	AB83
	NC(=[NH2+])SCc1ccccc1Cl	Q57	CHEMBL1229095
	NC(=[NH2+])SCc1ccc(Cl)cc1	Q58	CHEMBL1229097
	CC[NH+](CC)c1ccc(cc1)C(=O)NN	Q59	CHEMBL1358859
	C([NH2+]c1ccc2OCCOc2c1)c1ccncc1	Q60	CHEMBL1533396

	O=C(Cc1ccccc1)Nn1c[nH+]c2ccccc12	Q61	CHEMBL1560712
	NC1=[NH+]Cc2ccccc12	Q62	CHEMBL1617729
	O=C(C[NH+]1CCCCC1)Nc1ccc2OCOc2c1	Q63	CHEMBL1921970
	C[NH2+]Cc1ccc(Oc2cccnc2)o1	Q64	CHEMBL1921971
	NC(=[NH2+])c1ccc(cc1)C(F)(F)F	Q65	CHEMBL1921972
	O=C(Cc1cn2ccccc2[nH+]1)Nc1ccccc1	Q66	CHEMBL1921973

	OC(COCc1ccccc1F)C[NH+]1CCCC1	Q67	CHEMBL1921974

PrenDB: A substrate prediction database to enable biocatalytic use of prenyltransferases

Jakub Gunera[‡,†,1], Florian Kindinger[§,1], Shu-Ming Li[§,†] and Peter Kolb[‡,†]

[‡]Department of Pharmaceutical Chemistry, Philipps-University, Marburg, Hesse, 35032, Germany

[§]Department of Pharmaceutical Biology and Biotechnology, Philipps-University, Marburg, Hesse, 35032, Germany

[†]Synmikro, LOEWE Centre for Synthetic Microbiology, Philipps-University, Marburg, Hesse, 35043, Germany

[1]These authors contributed equally to this study.

1. Substrates within PrenDB

Table S1: List of substrates with their PrenDB ID, SMILES string and cluster membership (C)

ID	PrenDB ID	SMILES	C
S1	PTDBSUB00695	O=C1NC(Cc2c[nH]c3ccccc23)C(=O)N2CCC[C@@H]12	1
S2	PTDBSUB00027	O=C1N[C@H](Cc2c[nH]c3ccccc23)C(=O)N2CCC[C@H]12	1
S3	PTDBSUB00022	O=C1N[C@H](Cc2c[nH]c3ccccc23)C(=O)N2CCC[C@@H]12	1
S4	PTDBSUB00001	O=C1N[C@@H](Cc2c[nH]c3ccccc23)C(=O)N2CCC[C@@H]12	1
S5	PTDBSUB00017	O=C1N[C@@H](Cc2c[nH]c3ccccc23)C(=O)N2CCC[C@H]12	1
S6	PTDBSUB00286	O=C1N[C@H](Cc2c[nH]c3ccccc23)C(=O)Nc2ccccc21	2
S7	PTDBSUB00392	O=C1N[C@@H](Cc2c[nH]c3ccccc23)C(=O)Nc2ccccc21	2
S8	PTDBSUB00049	C[C@@H]1NC(=O)[C@@H](Cc2c[nH]c3ccccc23)NC1=O	2
S9	PTDBSUB00044	C[C@H]1NC(=O)[C@@H](Cc2c[nH]c3ccccc23)NC1=O	2
S10	PTDBSUB00032	C[C@@H]1NC(=O)[C@H](Cc2c[nH]c3ccccc23)NC1=O	2
S11	PTDBSUB00039	C[C@H]1NC(=O)[C@H](Cc2c[nH]c3ccccc23)NC1=O	2
S12	PTDBSUB00012	O=C1N[C@@H](Cc2c[nH]c3ccccc23)C(=O)N[C@H]1Cc1ccccc1	2
S13	PTDBSUB00005	O=C1N[C@@H](Cc2c[nH]c3ccccc23)C(=O)N[C@H]1Cc1ccc(O)cc1	2
S14	PTDBSUB00412	O=C1N[C@H](Cc2c[nH]c3ccccc23)C(=O)N[C@H]1Cc1ccc(O)cc1	2
S15	PTDBSUB00059	O=C1CNC(=O)[C@H](Cc2c[nH]c3ccccc23)N1	2
S16	PTDBSUB00072	O=C1N[C@@H](Cc2c[nH]c3ccccc23)C(=O)N[C@H]1Cc1c[nH]c2ccccc12	2
S17	PTDBSUB00054	O=C1N[C@@H](Cc2c[nH]c3ccccc23)C(=O)N[C@H]1Cc1cnc[nH]1	2
S18	PTDBSUB00064	CC(C)C[C@@H]1NC(=O)[C@H](Cc2c[nH]c3ccccc23)NC1=O	2
S19	PTDBSUB00103	[NH3+][C@@H](Cc1c[nH]c2ccccc12)C(=O)[O-]	3
S20	PTDBSUB00173	[NH3+][C@H](Cc1c[nH]c2ccccc12)C(=O)[O-]	3
S21	PTDBSUB00186	C[C@@]([NH3+])(Cc1c[nH]c2ccccc12)C(=O)[O-]	3
S22	PTDBSUB00189	C[C@]([NH3+])(Cc1c[nH]c2ccccc12)C(=O)[O-]	3
S23	PTDBSUB00179	[NH3+]CCc1c[nH]c2ccccc12	3
S24	PTDBSUB00415	O=C([O-])Cc1c[nH]c2ccccc12	3
S25	PTDBSUB00491	O=C([O-])C(=O)Cc1c[nH]c2ccccc12	3
S26	PTDBSUB00297	O=C([O-])CCc1c[nH]c2ccccc12	3

S27	PTDBSUB00496	O=C([O-])CCCc1c[nH]c2ccccc12	3
S28	PTDBSUB00676	CC(=O)N[C@@H](Cc1c[nH]c2ccccc12)C(C)=O	3
S29	PTDBSUB00317	CC(=O)N[C@@H](Cc1c[nH]c2ccccc12)C(=O)[O-]	3
S30	PTDBSUB00423	CC(=O)N[C@H](Cc1c[nH]c2ccccc12)C(=O)[O-]	3
S31	PTDBSUB00183	C[NH2+][C@@H](Cc1c[nH]c2ccccc12)C(=O)[O-]	3
S32	PTDBSUB00192	O=C([O-])[C@@H](O)Cc1c[nH]c2ccccc12	3
S33	PTDBSUB00195	O=C([O-])[C@H](O)Cc1c[nH]c2ccccc12	3
S34	PTDBSUB00176	[NH3+][C@H](CC(=O)[O-])Cc1c[nH]c2ccccc12	3
S35	PTDBSUB00418	[NH3+][C@@H](CC(=O)[O-])Cc1c[nH]c2ccccc12	3
S36	PTDBSUB00139	[NH3+][C@@H](Cc1c[nH]c2ccccc12)C(=O)NCC(=O)[O-]	3
S37	PTDBSUB00488	[NH3+][C@H](Cc1c[nH]c2ccccc12)C(=O)NO	3
S38	PTDBSUB00499	COC(=O)[C@H]([NH3+])Cc1c[nH]c2ccccc12	3
S39	PTDBSUB00593	[NH3+][C@@H](C(=O)[O-])[C@@H](O)c1ccc(O)c(O)c1	4
S40	PTDBSUB00722	[NH3+][C@@H](C(=O)[O-])C(O)c1ccc(O)c(O)c1	4
S41	PTDBSUB00689	Cc1ccc(C[C@@H]2NC(=O)[C@H](Cc3ccc(O)cc3)NC2=O)cc1	5
S42	PTDBSUB00571	O=C([O-])[C@@H](O)Cc1ccc(O)cc1	5
S43	PTDBSUB00574	O=C([O-])[C@H](O)Cc1ccc(O)cc1	5
S44	PTDBSUB00554	[NH3+][C@@H](C(=O)[O-])c1ccc(O)cc1	5
S45	PTDBSUB00551	[NH3+][C@@H](CC(=O)[O-])c1ccc(O)cc1	5
S46	PTDBSUB00718	[NH3+]C(CC(=O)[O-])c1ccc(O)cc1	5
S47	PTDBSUB00376	CC([NH3+])(Cc1ccc(O)cc1)C(=O)[O-]	5
S48	PTDBSUB00578	C[C@@]([NH3+])(Cc1ccc(O)cc1)C(=O)[O-]	5
S49	PTDBSUB00467	O=C(CCc1ccc(O)cc1)c1c(O)cc(O)cc1O	5
S50	PTDBSUB00079	O=C([O-])C(=O)Cc1ccc(O)cc1	5
S51	PTDBSUB00568	O=C([O-])CCc1ccc(O)cc1	5
S52	PTDBSUB00666	Oc1cc2ccccc2cc1O	6
S53	PTDBSUB00652	Oc1ccc2ccccc2c1	6
S54	PTDBSUB00289	Oc1ccc2ccc(O)cc2c1	6
S55	PTDBSUB00613	Oc1ccc2cc(O)ccc2c1	6
S56	PTDBSUB00647	Nc1cccc2ccc(O)cc12	6
S57	PTDBSUB00448	COc1cccc2ccc(O)cc12	6
S58	PTDBSUB00451	CCOc1cccc2ccc(O)cc12	6
S59	PTDBSUB00656	Oc1ccc(O)c2ccccc12	6
S60	PTDBSUB00661	Oc1cccc2c(O)cccc12	6
S61	PTDBSUB00399	Oc1ccc2c(O)cccc2c1	6
S62	PTDBSUB00596	Oc1cccc2ccccc12	6
S63	PTDBSUB00445	Cc1ccc2cccc(O)c2c1	6
S64	PTDBSUB00601	Oc1ccc2cccc(O)c2c1	6
S65	PTDBSUB00088	COc1ccc2c3c4n(c2c1)[C@@H](C=C(C)C)OOC(C)(C)C[C@@H]4N1C(=O)[C@@H]2CCCN2C(=O)[C@]1(O)[C@H]3O	6
S66	PTDBSUB00082	COc1ccc2c3c([nH]c2c1)[C@H](C=C(C)C)N1C(=O)[C@@H]2CCCN2C(=O)[C@]1(O)[C@H]3O	6
S67	PTDBSUB00085	COc1ccc2c3c(n(CC=C(C)C)c2c1)[C@H](C=C(C)C)N1C(=O)[C@@H]2CCCN2C(=O)[C@]1(O)[C@H]3O	6
S68	PTDBSUB00094	Cc1cc2oc3cccc(=O)c-3c([O-])c2cc1O	6
S69	PTDBSUB00100	Cc1c([O-])c(C)c2c(O)c3c(=O)cccc-3oc2c1C	6
S70	PTDBSUB00091	Cc1cc2oc3cccc(=O)c-3c(O)c2c(CO)c1[O-]	6

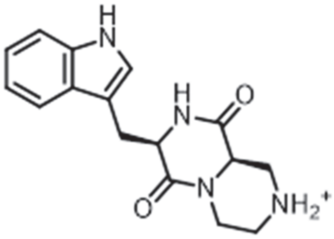
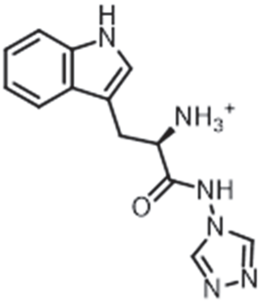
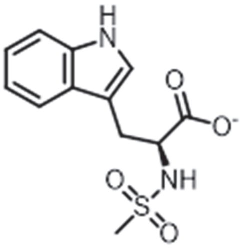
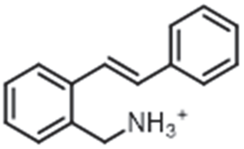
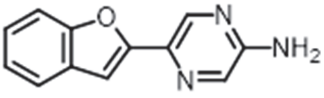
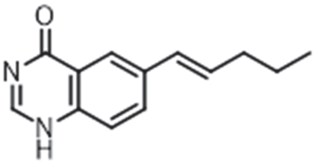
S71	PTDBSUB00097	Cc1cc2oc3cccc(=O)c-3c(O)c2c(C)c1[O-]	6
S72	PTDBSUB00479	COc1ccc(-c2coc3cc(=O)cc(O)c-3c2[O-])cc1	6
S73	PTDBSUB00472	O=c1cc2oc(-c3ccc(O)cc3)cc([O-])c-2c(O)c1	6
S74	PTDBSUB00343	O=c1cc2oc3cc(O)cc([O-])c3c([O-])c-2c(O)c1	6
S75	PTDBSUB00339	O=c1cc2oc3ccc(O)cc3c([O-])c-2c(O)c1	6
S76	PTDBSUB00351	O=c1cc2oc3cc(O)ccc3c([O-])c-2c(O)c1	6
S77	PTDBSUB00292	O=C1C=C([O-])c2c([O-])cc(O)cc2C1=O	6
S78	PTDBSUB00606	O=C(O)c1cc2cccc([O-])c2cc1O	6
S79	PTDBSUB00610	O=C(O)c1cc2cc(O)ccc2cc1[O-]	6
S80	PTDBSUB00483	O=C1C[C@H](c2ccccc2)Oc2cc(O)ccc21	6
S81	PTDBSUB00454	Cc1cc([O-])c2c(c1)O[C@H](c1ccc(O)cc1)CC2=O	6
S82	PTDBSUB00460	COc1ccc([C@@H]2CC(=O)c3c([O-])cc(C)cc3O2)cc1O	6
S83	PTDBSUB00464	COc1cc([C@H]2Oc3cc([C@H]4Oc5cc(O)cc([O-])c5C(=O)[C@@H]4O)ccc3O[C@@H]2CO)ccc1O	6
S84	PTDBSUB00457	Cc1cc(O)c2c(c1)O[C@H](c1ccc(O)c(O)c1)CC2O	6
S85	PTDBSUB00475	Oc1ccc(C2=COc3cc(O)cc(O)c3C2O)cc1	6
S86	PTDBSUB00367	Cn1c(O)c2c3c(c4c(c2c1O)=c1ccccc1=N4)N=c1ccccc1=3	6
S87	PTDBSUB00144	COc1c(O)c(=C2C=Nc3ccccc32)c(OC)c(O)c1=C1C=Nc2ccccc21	6
S88	PTDBSUB00396	O=C([O-])c1cccc2c1Nc1ccccc1N2	6
S89	PTDBSUB00385	CC(=O)O[C@H]1[C@@H](C)C[NH+](C)[C@@H]2Cc3c[nH]c4cccc(c34)[C@@H]12	7
S90	PTDBSUB00388	CC(=O)O[C@H]1[C@H](C)C[NH+](C)[C@@H]2Cc3c[nH]c4cccc(c34)[C@@H]12	7
S91	PTDBSUB00136	CC(=O)/C=C/c1c[nH]c2ccccc12	8
S92	PTDBSUB00300	O=C([O-])/C=C/c1c[nH]c2ccccc12	8
S93	PTDBSUB00359	O=C1NCc2c1c1c3ccccc3[nH]c1c1[nH]c3ccccc3c21	8
S94	PTDBSUB00364	O=C1N[C@@H](O)c2c1c1c3ccccc3[nH]c1c1[nH]c3ccccc3c21	8
S95	PTDBSUB00148	C=c1c(OC)c(-c2c(CC=C(C)C)[nH]c3ccccc23)c(=C)c(OC)c1-c1c[nH]c2ccccc12	8
S96	PTDBSUB00151	C=CC(C)(C)n1cc(C2=C(OC)C(=O)C(c3c[nH]c4ccccc34)=C(OC)C2=O)c2ccccc21	8
S97	PTDBSUB00703	[NH3+][C@H](Cc1cc(I)c([O-])c(I)c1)C(=O)[O-]	9
S98	PTDBSUB00370	[NH3+][C@@H](Cc1cc(Br)c([O-])c(Br)c1)C(=O)[O-]	9
S99	PTDBSUB00373	[NH3+][C@@H](Cc1cc(I)c([O-])c(I)c1)C(=O)[O-]	9
S100	PTDBSUB00548	[NH3+][C@@H](Cc1ccc([O-])c([N+](=O)[O-])c1)C(=O)[O-]	9
S101	PTDBSUB00706	[NH3+][C@H](Cc1ccc([O-])c([N+](=O)[O-])c1)C(=O)[O-]	9
S102	PTDBSUB00106	[NH3+][C@@H](Cc1ccc(O)cc1)C(=O)[O-]	9
S103	PTDBSUB00542	[NH3+][C@H](Cc1ccc(O)cc1)C(=O)[O-]	9
S104	PTDBSUB00321	Nc1ccc(C[C@H]([NH3+])C(=O)[O-])cc1	9
S105	PTDBSUB00545	Nc1ccc(C[C@@H]([NH3+])C(=O)[O-])cc1	9
S106	PTDBSUB00587	[NH3+][C@H](Cc1ccc(O)c(I)c1)C(=O)[O-]	9
S107	PTDBSUB00584	[NH3+][C@@H](Cc1ccc(O)c(F)c1)C(=O)[O-]	9
S108	PTDBSUB00382	[NH3+][C@@H](Cc1ccc(O)c(I)c1)C(=O)[O-]	9
S109	PTDBSUB00581	[NH3+][C@H](Cc1ccc(O)c(F)c1)C(=O)[O-]	9
S110	PTDBSUB00112	[NH3+][C@@H](Cc1cccc(O)c1)C(=O)[O-]	9
S111	PTDBSUB00565	[NH3+][C@H](Cc1cccc(O)c1)C(=O)[O-]	9
S112	PTDBSUB00562	Nc1cc(C[C@@H]([NH3+])C(=O)[O-])ccc1O	9
S113	PTDBSUB00379	[NH3+][C@@H](Cc1ccc(O)c(O)c1)C(=O)[O-]	9
S114	PTDBSUB00590	[NH3+][C@H](Cc1ccc(O)c(O)c1)C(=O)[O-]	9

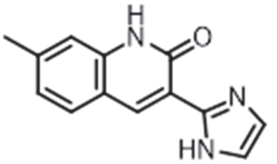
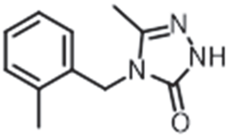
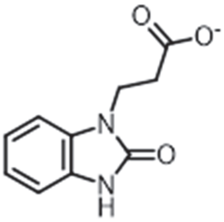
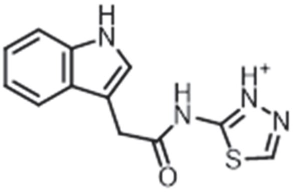
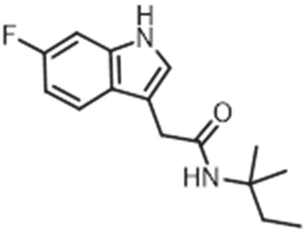
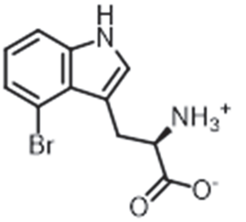
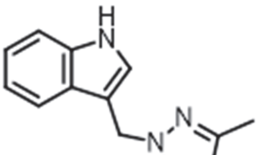
S115	PTDBSUB00161	Cc1ccc2c(C[C@H]([NH3+])C(=O)[O-])c[nH]c2c1	10
S116	PTDBSUB00164	Cc1ccc2c(C[C@@H]([NH3+])C(=O)[O-])c[nH]c2c1	10
S117	PTDBSUB00310	[NH3+][C@H](Cc1c[nH]c2cc(F)ccc12)C(=O)[O-]	10
S118	PTDBSUB00441	[NH3+][C@@H](Cc1c[nH]c2cc(F)ccc12)C(=O)[O-]	10
S119	PTDBSUB00516	[NH3+][C@H](Cc1c[nH]c2ccc(F)cc12)C(=O)[O-]	10
S120	PTDBSUB00438	[NH3+][C@@H](Cc1c[nH]c2ccc(F)cc12)C(=O)[O-]	10
S121	PTDBSUB00432	[NH3+][C@@H](Cc1c[nH]c2ccc(Br)cc12)C(=O)[O-]	10
S122	PTDBSUB00435	[NH3+][C@H](Cc1c[nH]c2ccc(Br)cc12)C(=O)[O-]	10
S123	PTDBSUB00167	COc1ccc2[nH]cc(C[C@H]([NH3+])C(=O)[O-])c2c1	10
S124	PTDBSUB00170	COc1ccc2[nH]cc(C[C@@H]([NH3+])C(=O)[O-])c2c1	10
S125	PTDBSUB00155	Cc1ccc2[nH]cc(C[C@H]([NH3+])C(=O)[O-])c2c1	10
S126	PTDBSUB00158	Cc1ccc2[nH]cc(C[C@@H]([NH3+])C(=O)[O-])c2c1	10
S127	PTDBSUB00429	[NH3+][C@@H](Cc1c[nH]c2ccc(O)cc12)C(=O)[O-]	10
S128	PTDBSUB00511	[NH3+][C@H](Cc1c[nH]c2ccc(O)cc12)C(=O)[O-]	10
S129	PTDBSUB00303	Cn1cc(C[C@@H]([NH3+])C(=O)[O-])c2ccccc21	11
S130	PTDBSUB00109	[NH3+][C@@H](Cc1ccccc1O)C(=O)[O-]	11
S131	PTDBSUB00558	[NH3+][C@H](Cc1ccccc1O)C(=O)[O-]	11
S132	PTDBSUB00313	Cc1cccc2c(C[C@@H]([NH3+])C(=O)[O-])c[nH]c12	11
S133	PTDBSUB00407	Cc1cccc2c(C[C@H]([NH3+])C(=O)[O-])c[nH]c12	11
S134	PTDBSUB00120	Cc1cccc2[nH]cc(C[C@H]([NH3+])C(=O)[O-])c12	11
S135	PTDBSUB00306	Cc1cccc2[nH]cc(C[C@@H]([NH3+])C(=O)[O-])c12	11
S136	PTDBSUB00123	COc1cccc2[nH]cc(C[C@H]([NH3+])C(=O)[O-])c12	11
S137	PTDBSUB00126	Nc1cccc2[nH]cc(C[C@H]([NH3+])C(=O)[O-])c12	11
S138	PTDBSUB00274	CC(C)=CCc1cc(C)cc2[nH]cc(C[C@H]([NH3+])C(=O)[O-])c12	12
S139	PTDBSUB00276	CC(C)=CCc1cc(C)cc2[nH]cc(C[C@@H]([NH3+])C(=O)[O-])c12	12
S140	PTDBSUB00270	CC(C)=CCc1c(C)ccc2[nH]cc(C[C@H]([NH3+])C(=O)[O-])c12	12
S141	PTDBSUB00272	CC(C)=CCc1c(C)ccc2[nH]cc(C[C@@H]([NH3+])C(=O)[O-])c12	12
S142	PTDBSUB00278	COc1ccc2[nH]cc(C[C@H]([NH3+])C(=O)[O-])c2c1CC=C(C)C	12
S143	PTDBSUB00281	COc1ccc2[nH]cc(C[C@@H]([NH3+])C(=O)[O-])c2c1CC=C(C)C	12
S144	PTDBSUB00239	CC(C)=CCc1c(C)ccc2c(C[C@H]([NH3+])C(=O)[O-])c[nH]c12	12
S145	PTDBSUB00242	CC(C)=CCc1c(C)ccc2c(C[C@@H]([NH3+])C(=O)[O-])c[nH]c12	12
S146	PTDBSUB00233	CC(C)=CCc1cc(C)cc2c(C[C@H]([NH3+])C(=O)[O-])c[nH]c12	12
S147	PTDBSUB00236	CC(C)=CCc1cc(C)cc2c(C[C@@H]([NH3+])C(=O)[O-])c[nH]c12	12
S148	PTDBSUB00245	COc1cc(CC=C(C)C)c2[nH]cc(C[C@H]([NH3+])C(=O)[O-])c2c1	12
S149	PTDBSUB00248	COc1cc(CC=C(C)C)c2[nH]cc(C[C@@H]([NH3+])C(=O)[O-])c2c1	12
S150	PTDBSUB00251	CC(C)=CCc1cccc2[nH]cc(C[C@H]([NH3+])C(=O)[O-])c12	13
S151	PTDBSUB00253	CC(C)=CCc1cccc2[nH]cc(C[C@@H]([NH3+])C(=O)[O-])c12	13
S152	PTDBSUB00266	CC(C)=CCc1cccc2[nH]cc(C[C@H](O)C(=O)[O-])c12	13
S153	PTDBSUB00268	CC(C)=CCc1cccc2[nH]cc(C[C@@H](O)C(=O)[O-])c12	13
S154	PTDBSUB00255	CC(C)=CCc1cccc2[nH]cc(C[C@H]([NH3+])CC(=O)[O-])c12	13
S155	PTDBSUB00260	C[NH2+][C@@H](Cc1c[nH]c2cccc(CC=C(C)C)c12)C(=O)[O-]	13
S156	PTDBSUB00257	CC(C)=CCc1cccc2[nH]cc(CC[NH3+])c12	13
S157	PTDBSUB00262	CC(C)=CCc1cccc2[nH]cc(C[C@@](C)([NH3+])C(=O)[O-])c12	13
S158	PTDBSUB00264	CC(C)=CCc1cccc2[nH]cc(C[C@](C)([NH3+])C(=O)[O-])c12	13
S159	PTDBSUB00205	CC(C)=CCc1cccc2c(C[C@H]([NH3+])C(=O)[O-])c[nH]c12	14

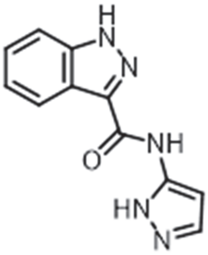
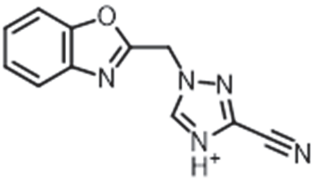
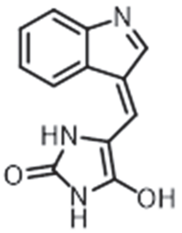
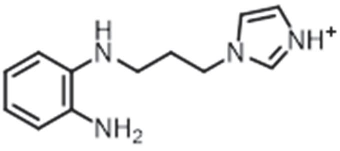
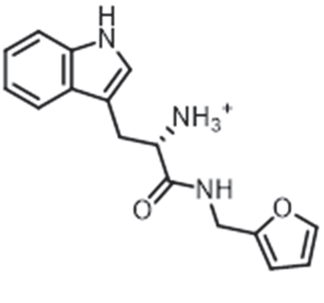
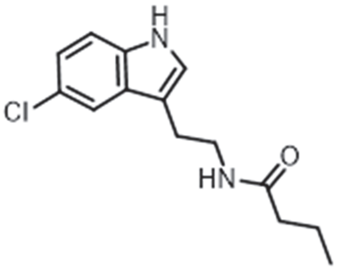
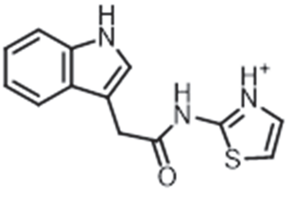
S160	PTDBSUB00208	CC(C)=CCc1cccc2c(C[C@@H]([NH3+])C(=O)[O-])c[nH]c12	14
S161	PTDBSUB00226	CC(C)=CCc1cccc2c(C[C@H](O)C(=O)[O-])c[nH]c12	14
S162	PTDBSUB00229	CC(C)=CCc1cccc2c(C[C@@H](O)C(=O)[O-])c[nH]c12	14
S163	PTDBSUB00211	CC(C)=CCc1cccc2c(C[C@H]([NH3+])CC(=O)[O-])c[nH]c12	14
S164	PTDBSUB00217	C[NH2+][C@@H](Cc1c[nH]c2c(CC=C(C)C)cccc12)C(=O)[O-]	14
S165	PTDBSUB00214	CC(C)=CCc1cccc2c(CC[NH3+])c[nH]c12	14
S166	PTDBSUB00220	CC(C)=CCc1cccc2c(C[C@@](C)([NH3+])C(=O)[O-])c[nH]c12	14
S167	PTDBSUB00223	CC(C)=CCc1cccc2c(C[C@](C)([NH3+])C(=O)[O-])c[nH]c12	14

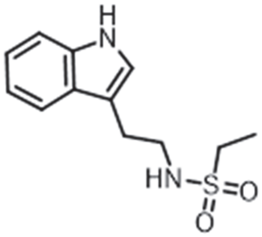
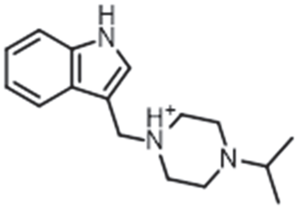
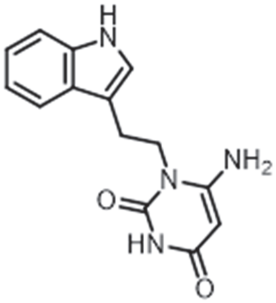
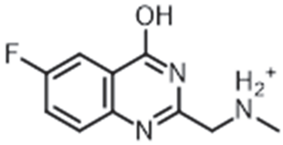
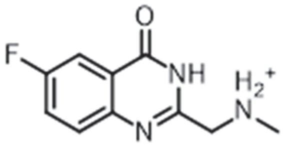
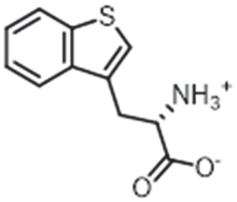
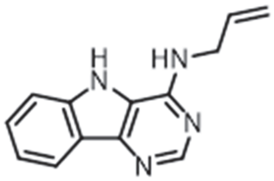
2. Virtual hits

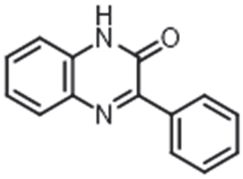
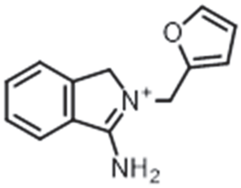
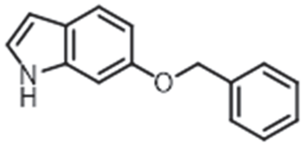
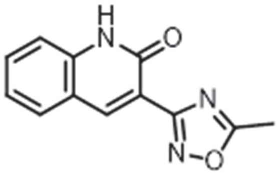
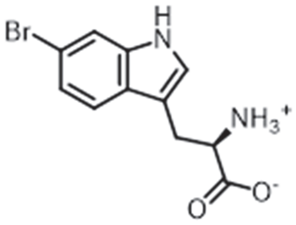
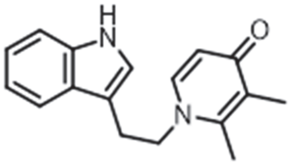
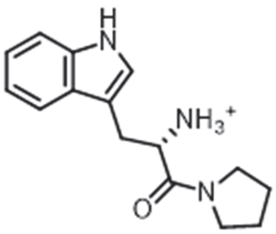
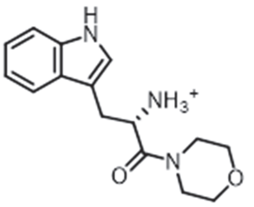
Table S2: Structures, IDs, enzyme-related yields and number of matched PrenDB reactions.

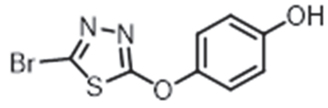
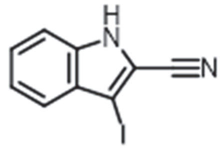
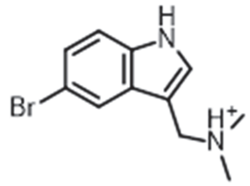
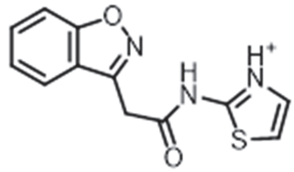
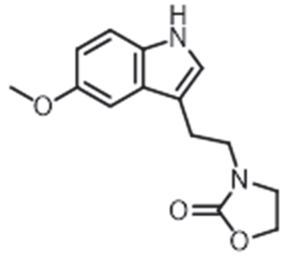
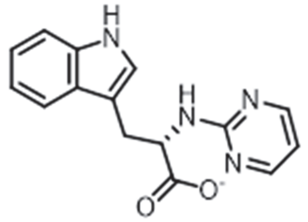
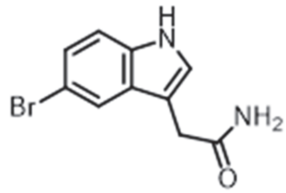
Substrate	ID	Product yield [%]			Matched reactions
		FtmPT1	FgaPT2	CdpNPT	
	1	94.1	18.2	41.3	299
	2	81.3	80.1	29.5	299
	3	-	3.2	22.1	299
	4	-	-	-	6
	5	-	-	0.9	33
	6	-	-	-	57

	7	7.0	-	10.1	57
	8	-	-	-	6
	9	-	5.3	-	17
	10	40.8	28.9	99.7	233
	11	48.1	17.1	88.1	238
	12	8.5	-	14.4	299
	13	71.0	31.7	47.2	231

	14	2.0	-	2.6	50
	15	-	-	-	58
	16	68.8	34.8	29.7	20
	17	-	-	-	14
	18	89.1	84.9	62.0	299
	19	76.4	6.2	46.9	233
	20	76.0	37.4	97.4	233

	21	98.6	99.2	99.3	233
	22	-	-	-	231
	23	64.8	75.0	65.2	233
	24[a]	-	-	-	50
	25	-	-	-	50
	26	-	-	-	33
	27	8.7	82.9	11.8	137

	28	-	-	-	23
	29[b]	-	-	-	6
	30	60.1	59.4	96.1	229
	31	-	-	-	50
	32	47.5	44.7	-	299
	33	99.5	99.1	87.2	233
	34	44.4	19.9	94.8	299
	35	35.5	16.3	94.1	299

	36[b]	-	-	-	77
	37	-	-	-	137
	38	-	-	-	235
	39[b]	3.8	-	2.8	33
	40	51.5	4.4	9.2	237
	41	9.9	22.5	13.5	325
	42	-	14.9	38.1	299

[a] Compound **24** is the imidic acid tautomer of **25** and was thus excluded from further consideration in this work. Total and relative numbers throughout the manuscript reflect the number of unique compounds, i.e. 38.

[b] These compounds could not be obtained as ordered and were excluded from further consideration in this work

Tryptophan *C5*-, *C6*- and *C7*-prenylating enzymes displaying a preference for *C6* of the indole ring in the presence of unnatural dimethylallyl diphosphate analogs

Julia Winkelblech[1,2,‡], Mike Liebhold[1,‡], Jakub Gunera[2,3], Xiulan Xie[4], Peter Kolb[2,3] and Shu-Ming Li[1,2]

[1]Department of Pharmaceutical Biology and Biotechnology, Philipps-University, Marburg, Hesse, 35032, Germany

[2]Synmikro, LOEWE Centre for Synthetic Microbiology, Philipps-University, Marburg, Hesse, 35043, Germany

[3]Department of Pharmaceutical Chemistry, Philipps-University, Marburg, Hesse, 35032, Germany

[4]Department of Chemistry, Philipps-University, Marburg, Hesse, 35032, Germany

[‡]These authors contributed equally to this study.

1. Structure elucidation

In the ^{1}H-NMR spectra of **Ia**, signals of the indole moiety at 7.60 (1H, dd, 8.1, 0.4), 7.15 (1H, d, 0.7), 7.12 (1H, s) and 6.89 ppm (1H, dd, 8.1, 1.4) superimposed with those for H-4, H-7, H-2 and H-5 of 6-methylallyl-l-tryptophan, respectively.[1] The signals of H-10 and H-11 as well as H-1' and H-4' of **Ia** were also overlapping almost completely (maximum shift 0.03 ppm) with those of 6-methylallyl-l-tryptophan.[1] This proved unequivocally the regular *C6*-alkylation of l-tryptophan with MAPP as alkyl donor in the presence of the tested prenyltransferases (6-DMATS$_{Sa}$, 6-DMATS$_{Sv}$, TyrPT, 5-DMATS$_{Sc}$ and 5-DMATS). Comparing the ^{1}H-NMR spectrum of **IIa** with that of C6-(2-pentenyl)-l-tryptophan[1] showed nearly identical chemical shifts and coupling patterns for all of the protons. This verified the regular alkylation of l-tryptophan at position C-6 of the indole ring by using 2-pentenyl-PP as alkyl donor. The aromatic protons of **IIIa** at 7.60 (1H, dd, 8.2, 0.6), 7.18 (1H, s), 7.12 (1H, s) and 6.93 ppm (1H, dd, 8.2, 1.5) showed the same coupling pattern and chemical shifts as observed for **Ia** and **IIa**. These signals also corresponded to those of *C6*-alkylated l-tryptophan derivatives[1] and therefore proved the *C6*-benzylation of l-tryptophan. The chemical shifts observed for H-10 and H-11 at 3.48, 3.11 and 3.82 ppm, also overlapped very well with those of the other C6-alkylated l-tryptophan derivatives. The five additional aromatic protons and two additional aliphatic protons observed in the ^{1}H-NMR spectra of **IIIa** confirmed the presence of the benzyl moiety.

From the incubation mixtures of TyrPT and 5-DMATS$_{Sc,}$ the regular alkylated products **Ib1** and **IIb1** as well as the regular benzylated product **IIIb** were isolated. The ^{1}H-NMR spectrum of **IIb1** showed one singlet at 7.19, two doublets at 7.55 and 6.93 and one triplet at 7.00 ppm for one proton each. This indicated an alkylation at position C-4 or C-7 of the indole ring. In the HMBC spectrum of **IIb1** (Figure S16-S19), correlations between H-10 at 3.13 and C-2 at 124.7 ppm, C-11 at 56.5 ppm with two quaternary carbon atoms at 109.9 and 128.4 ppm were observed. Correlations between the proton at 7.00 ppm, which is either H-5 or H-6, and two quaternary carbon atoms at 128.4 and at 125.1 ppm but not with that at 109.9 ppm were detected. Consequently, the quaternary carbons at 109.9 and 128.4 ppm were assigned to C-3 and C-9, respectively. Therefore, the signal at 7.00 ppm was assigned to H-5. Further correlations were found between the doublet at 7.56 ppm and the quaternary carbons C-3, C-9 and another one at 136.8 ppm. These correlations are only possible, if the proton at

7.56 ppm is for H-4 and the carbons at 136.8 and 125.1 ppm are for C-8 and C-7, respectively. Thus, an alkylation at position C-7 was proven. This was further confirmed by correlations between H-1' at 3.54 ppm and the quaternary carbons at 125.1 and 136.8 ppm, but not with that at 128.4 ppm (C-9). In addition, the doublet at 6.93 ppm also correlated with the quaternary carbons at 125.1 and 136.8 ppm as well as the signal for H-1'. The signals at 125.1, 136.8 and 6.93 ppm were assigned to C-7, C-8 and H-6, accordingly.

The chemical shifts of the protons in the tryptophan moiety of **Ib1** at 7.56 (H-4), 7.19 (H-2), 7.00 (H-5), 6.93 (H-6), 3.85 (H-11), 3.51 (H-10) and 3.14 ppm (H-10) almost completely overlapped with those of **IIb1**. Similar spectrum was obtained for **IIIb**. Therefore, the alkylation position in **Ib1** and benzylation position in **IIIb** were assigned unequivocally to C-7 of the indole ring.

In the presence of TyrPT, the additional products **Ib2** and **IIb2** were detected by using MAPP and 2-pentenyl-PP as alkyl donor, respectively. The coupling patterns and chemical shifts of the aromatic protons of both compounds corresponded very well to those of **Ib1** and **IIb1**, confirming a *C7*-alkylation of l-tryptophan (Slight shifts of approximate 0.05 ppm were observed). However, the signals of the alkyl residues of **Ib2** and **IIb2**, displayed distinct chemical shifts and coupling patterns in comparison to those of **Ib1** and **IIb1**. The coupling pattern for H-1' at 5.12 (1H, dt, 17.3, 1.7) and 5.04 ppm (1H, dt, 10.3, 1.7) as well as for H-2' at 6.12 ppm (1H, ddd, 17.3, 10.3, 6.3) in the spectrum of **Ib2** showed clearly a reverse alkylation.[2] The same was true for H-1' at 5.08 (1H, dt, 17.2, 1.5) and 4.99 ppm (1H, ddd, 10.2, 1.9, 1.0) as well as for H-2' at 6.06 ppm (1H, ddd, 17.2, 10.2, 7.6) in the spectrum of **IIb2**. This proved the reverse orientation of the alkyl residues of both compounds. Consequently, **Ib2** and **IIb2** were identified as 7-(3'-methylallyl-)-l-tryptophan and 7-(3'-pentenyl-)-l-tryptophan, respectively. From the ^{1}H NMR spectra of **Ib2** and **IIb2**, it was evident that only one of the two possible diastereomers was isolated. Unfortunately, the stereochemistry of these compounds at position C-3' could not be determined in this study.

Due to low conversion and unsuccessful separation on HPLC, **Ic**, **IIc** and **IIIc** were elucidated from the mixture with **Ia**, **IIa** and **IIIa**, respectively. The aliphatic signals of the indole moiety and those of the alkyl or benzyl residue for **Ic** were overlapped by those of **Ia**. The aromatic signals of the indole moiety were distinct from those of **Ia**, and could be used to identify the alkylation position. Comparison of the NMR data obtained in this study with those published previously [1,3] confirmed that **Ic**, **IIc** and **IIIc** to be 5-methylallyl-l-tryptophan, 5-(2-pentenyl)-l-tryptophan and 5-benzyl-l-tryptophan [3], respectively.

Reference List

1. Liebhold, M.; Xie, X.; Li, S.-M. *Org. Lett.* **2012,** *14*, 4884-4885.
2. Liebhold, M.; Xie, X.; Li, S.-M. *Org. Lett.* **2013,** *15*, 3062-3065.
3. Liebhold, M.; Li, S.-M. *Org. Lett.* **2013,** *15*, 5834-5837.

Table S1: Enzyme activities of several prenyltransferases toward l-tryptophan in the presence of DMAPP and its analogues MAPP, 2-pentenyl-PP and benzyl-PP.

	DMAPP [%]	2-pentenyl-PP [%]	MAPP [%]	benzyl-PP [%]
6-DMATS$_{Sa}$	99.9±0.2	91.2±0.07	51.1±0.5	13.9±0.3
6-DMATS$_{Sv}$	99.4±0.9	89.3±0.6	37.6±0.3	8.2±0.3
TyrPT	68.5±0.2	38.3±0.6	17.7±0.2	8.5±1.0
5-DMATS$_{Sc}$	81.4±2.4	65.0±0.1	21.0±0.4	6.6±0.1
5-DMATS	99.8±0.35	91.9±0.1	58.9±0.07	28.6±0.07

The reaction mixtures contained 1 mM L-tryptophan and 2mM DMAPP, 2-pentenyl-PP, MAPP or benzyl-PP and were incubated with 7.5 µM of purified protein at 37°C for 16 h. Conversion yields are given as mean of two independent measurements.

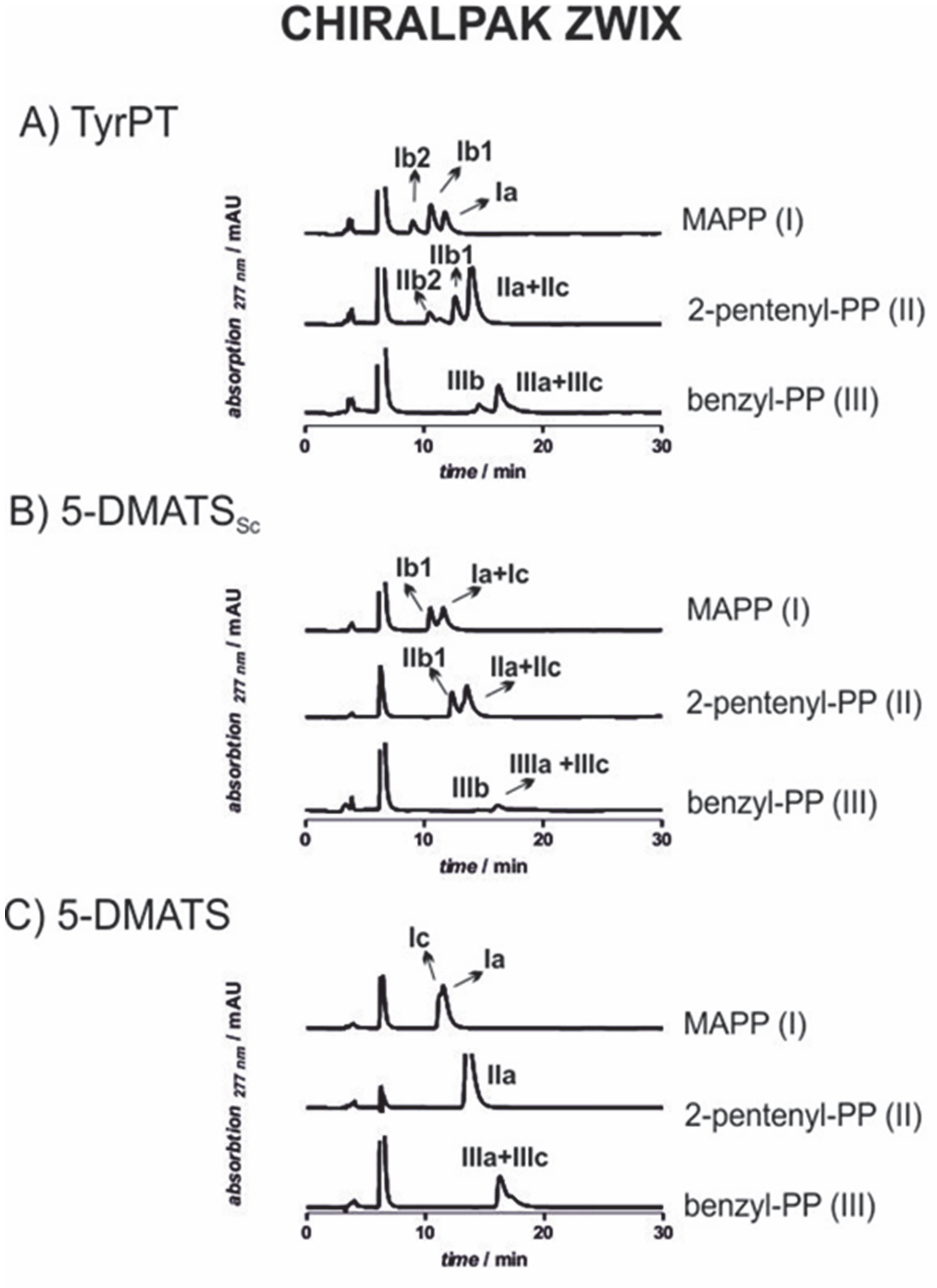


Figure S1: HPLC analysis of the reaction mixtures of L-tryptophan with MAPP (**I**), 2-pentenyl-PP (**II**) and benzyl-PP (**III**) on a Chiralpak Zwix (+) column. The enzyme assays of 100 µL contained 1 mM L-tryptophan, 2 mM alkyl or benzyl diphosphate, 5 mM $CaCl_2$ and 7.5 µM of purified protein were incubated at 37°C for 16 h.

6-DMATS$_{Sa}$

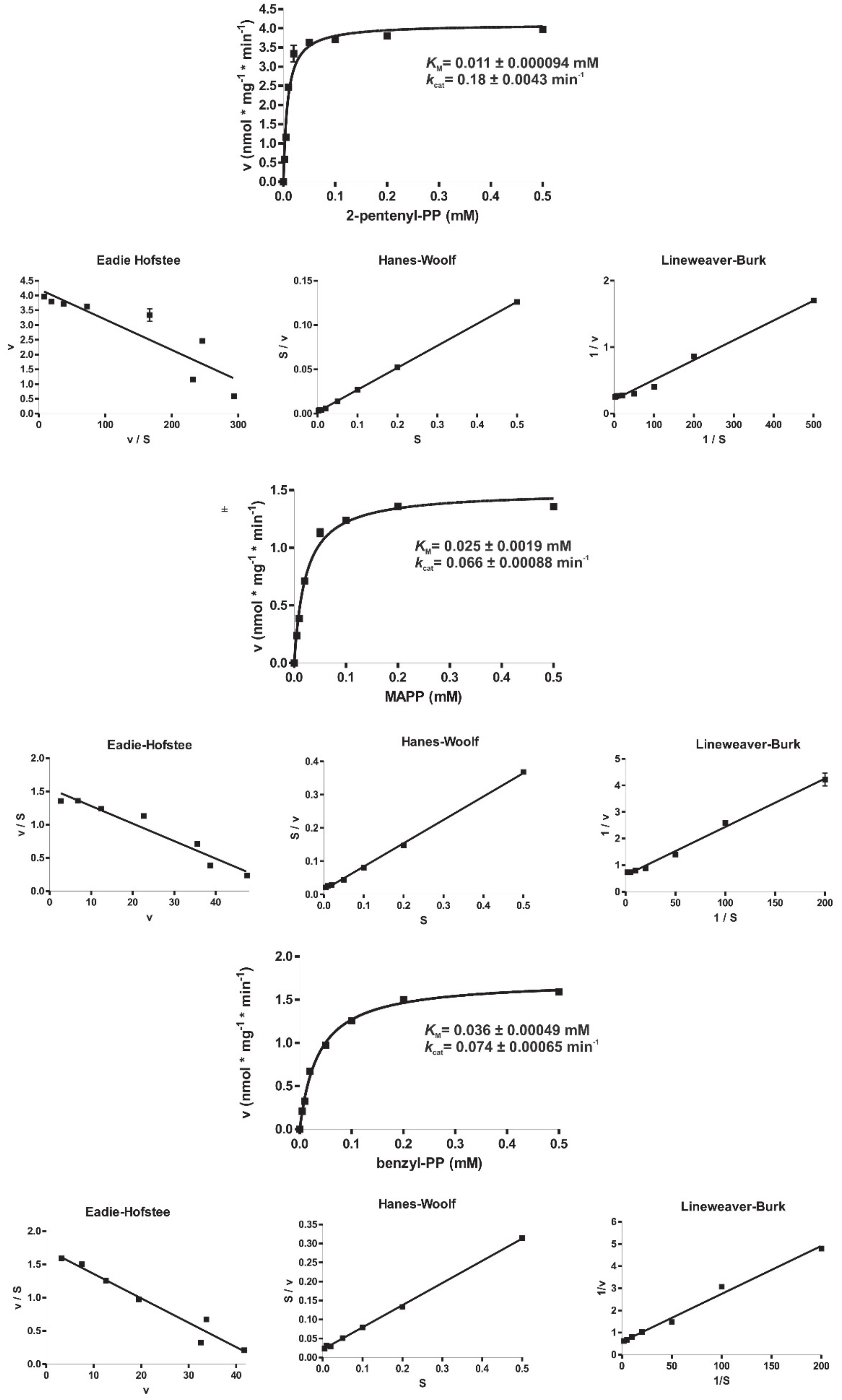


Figure S2: Dependence of the product formation of the 6-DMATS$_{Sa}$ reaction on the presence of 2-pentenyl-PP, methylallyl-PP (MAPP) or benzyl-PP with L-tryptophan.

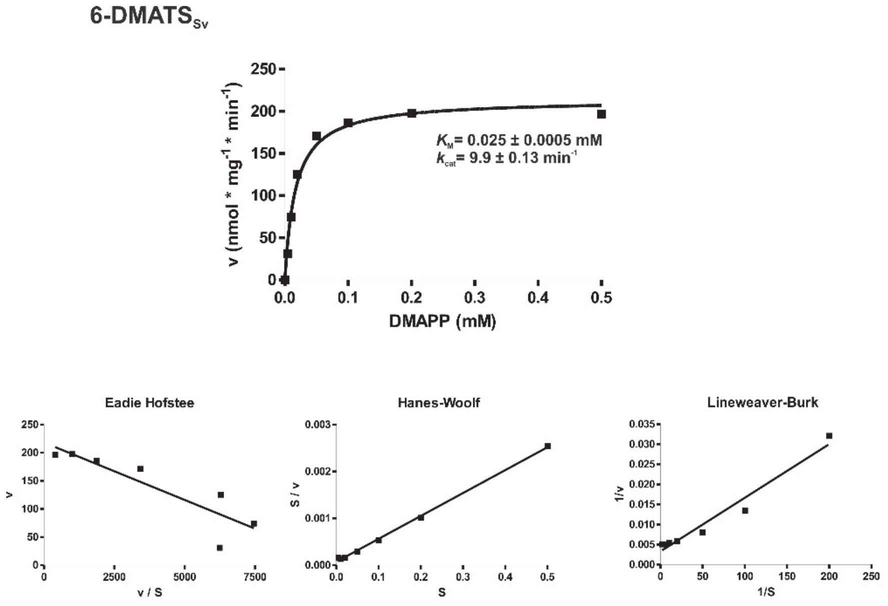


Figure S3: Dependence of the product formation of the 6-DMATS$_{Sv}$ reaction on the presence of DMAPP with L-tryptophan.

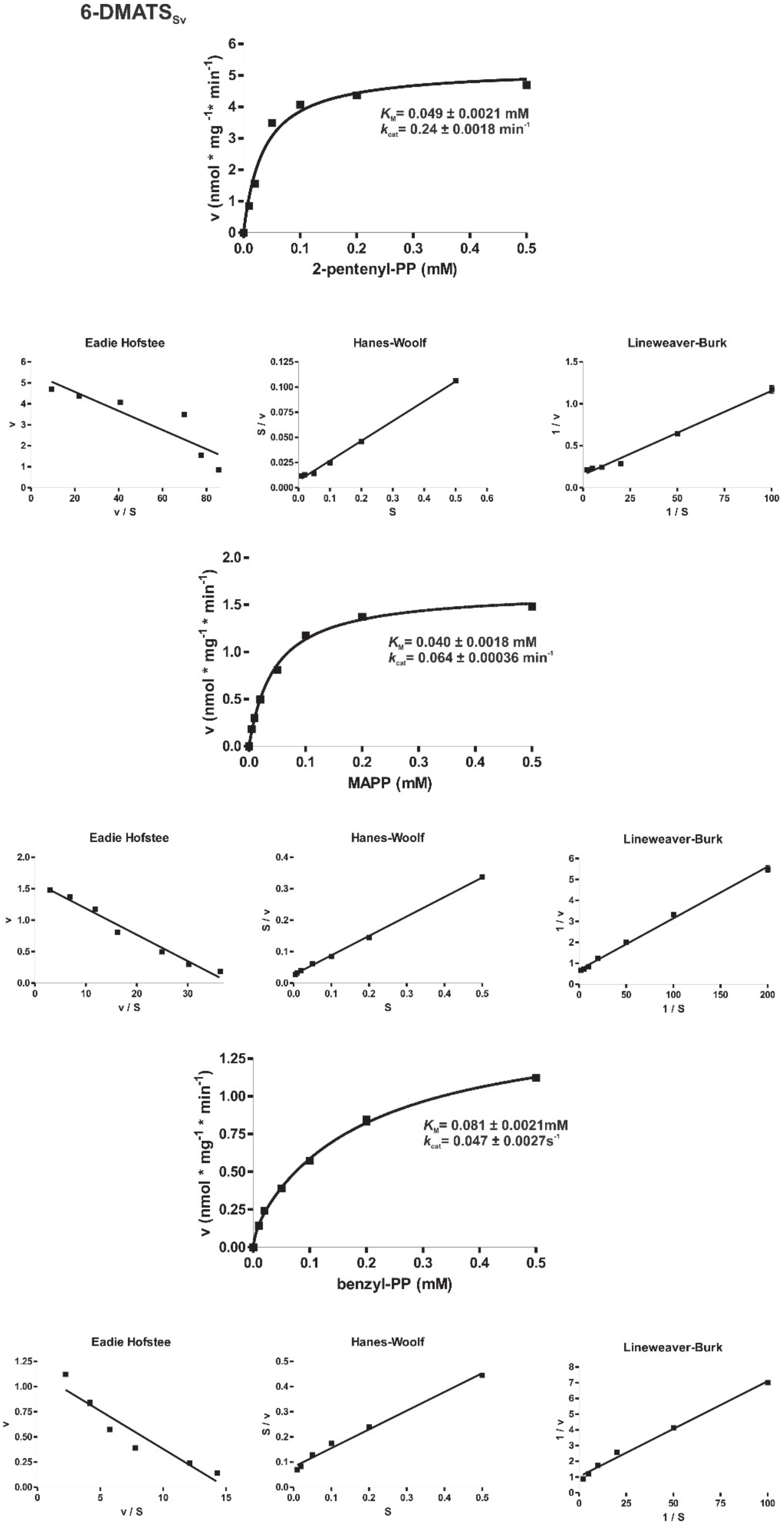


Figure S4: Dependence of the product formation of the 6-DMATS$_{Sv}$ reaction on the presence of 2-pentenyl-PP, methylallyl-PP (MAPP) or benzyl-PP with L-tryptophan.

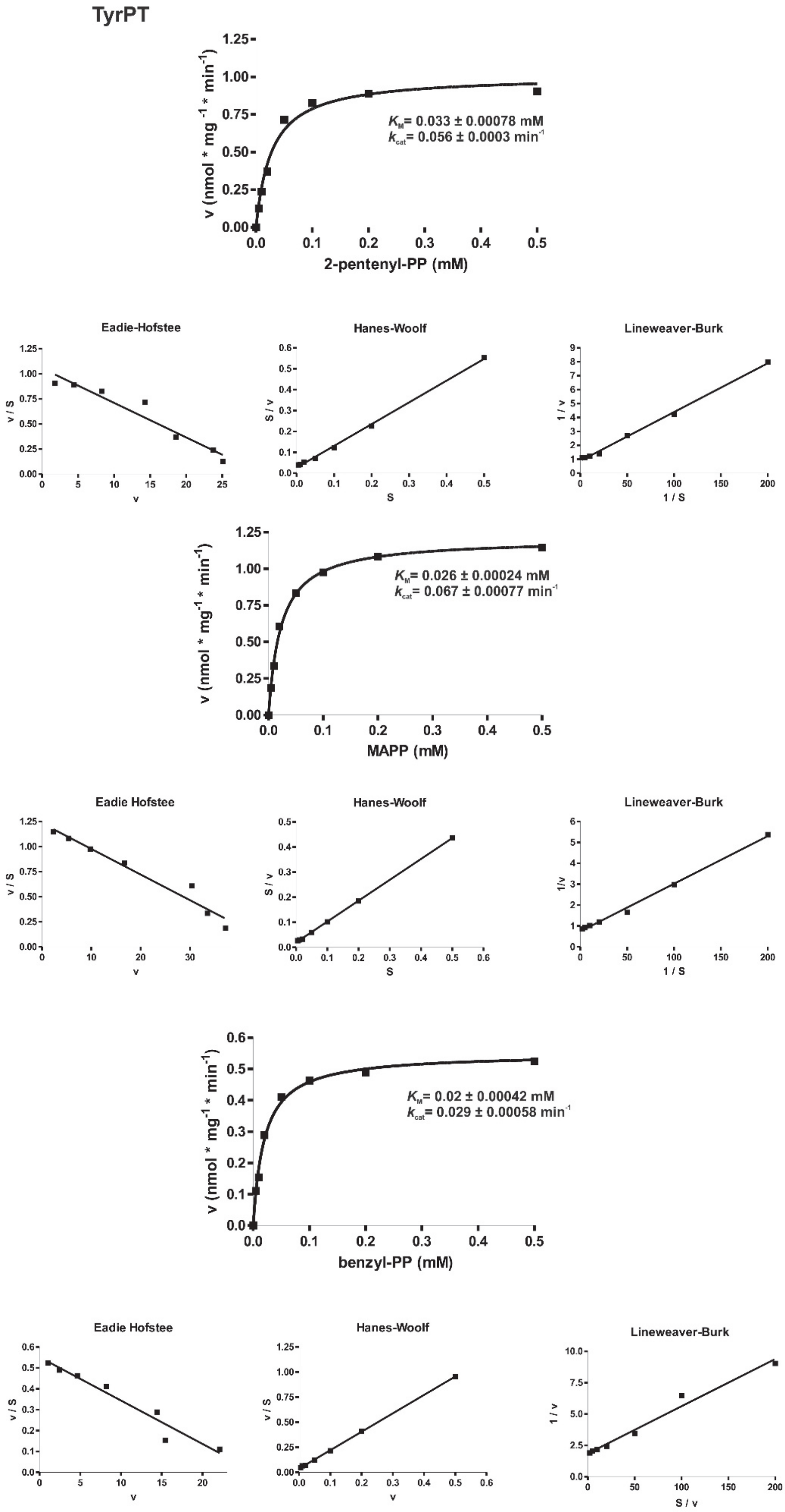


Figure S5: Dependence of the product formation of the TyrPT reaction on the presence of 2-pentenyl-PP, methylallyl-PP (MAPP) or benzyl-PP with L-tryptophan.

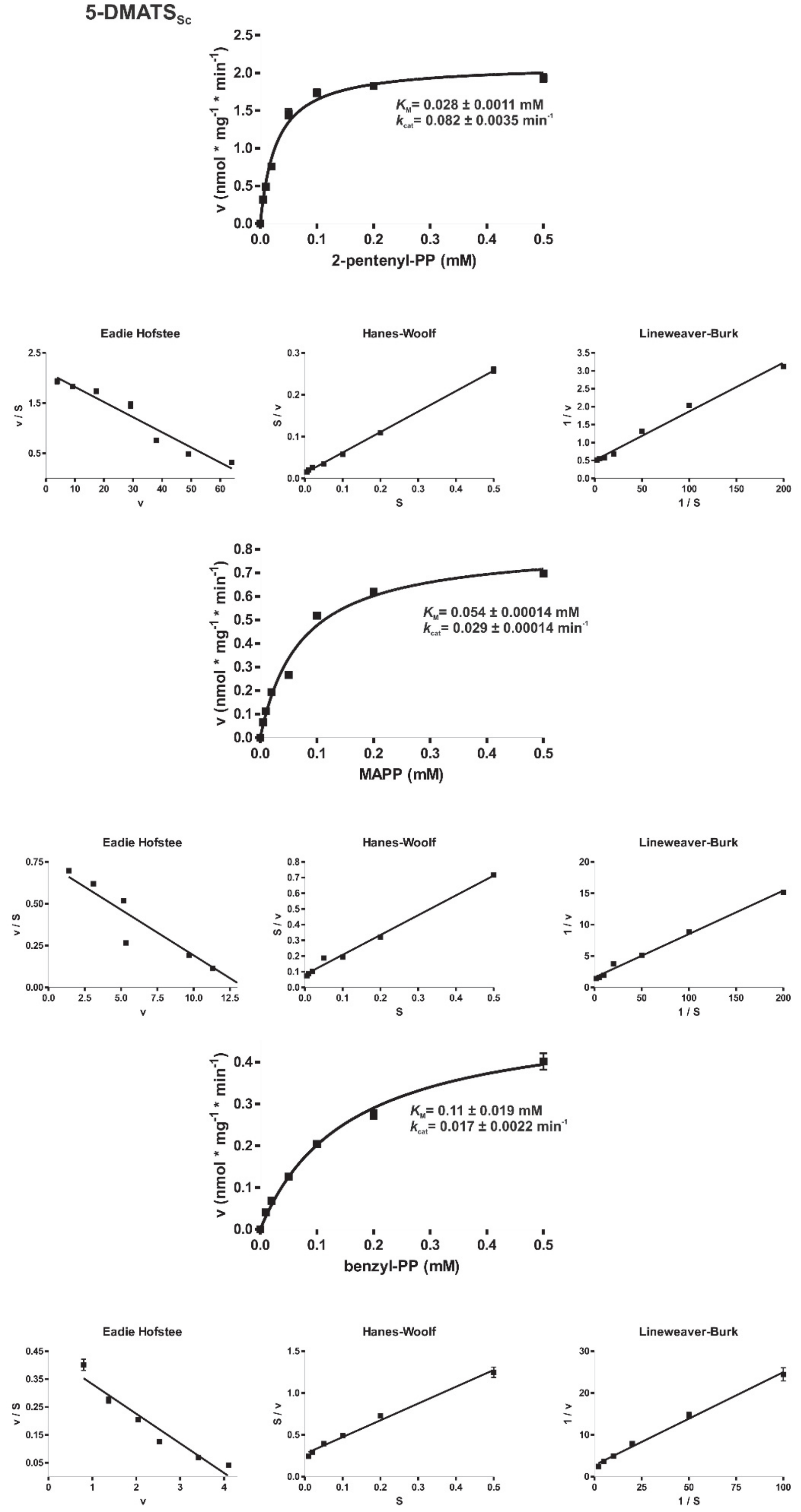


Figure S6: Dependence of the product formation of the 5-DMATS$_{Sc}$ reaction on the presence of 2-pentenyl-PP, methylallyl-PP (MAPP) or benzyl-PP with L-tryptophan.

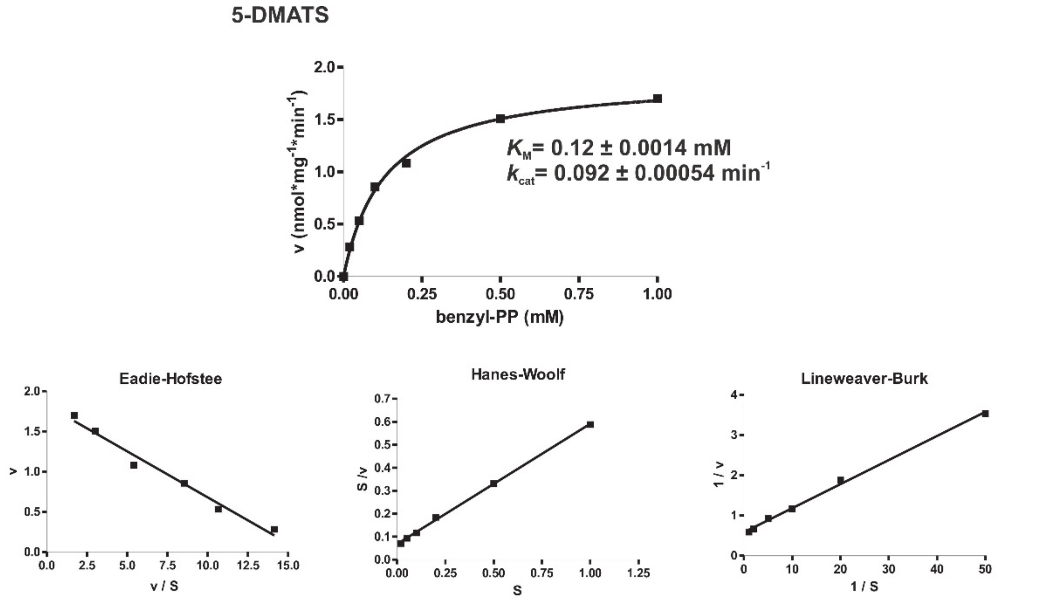


Figure S7: Dependence of the product formation of the 5-DMATS reaction on the presence of benzyl-PP with L-tryptophan.

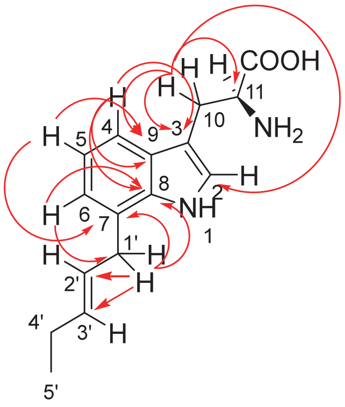


Figure S8: HMBC connectivities of 7-(2-pentenyl-)-L-tryptophan (**IIb1**).

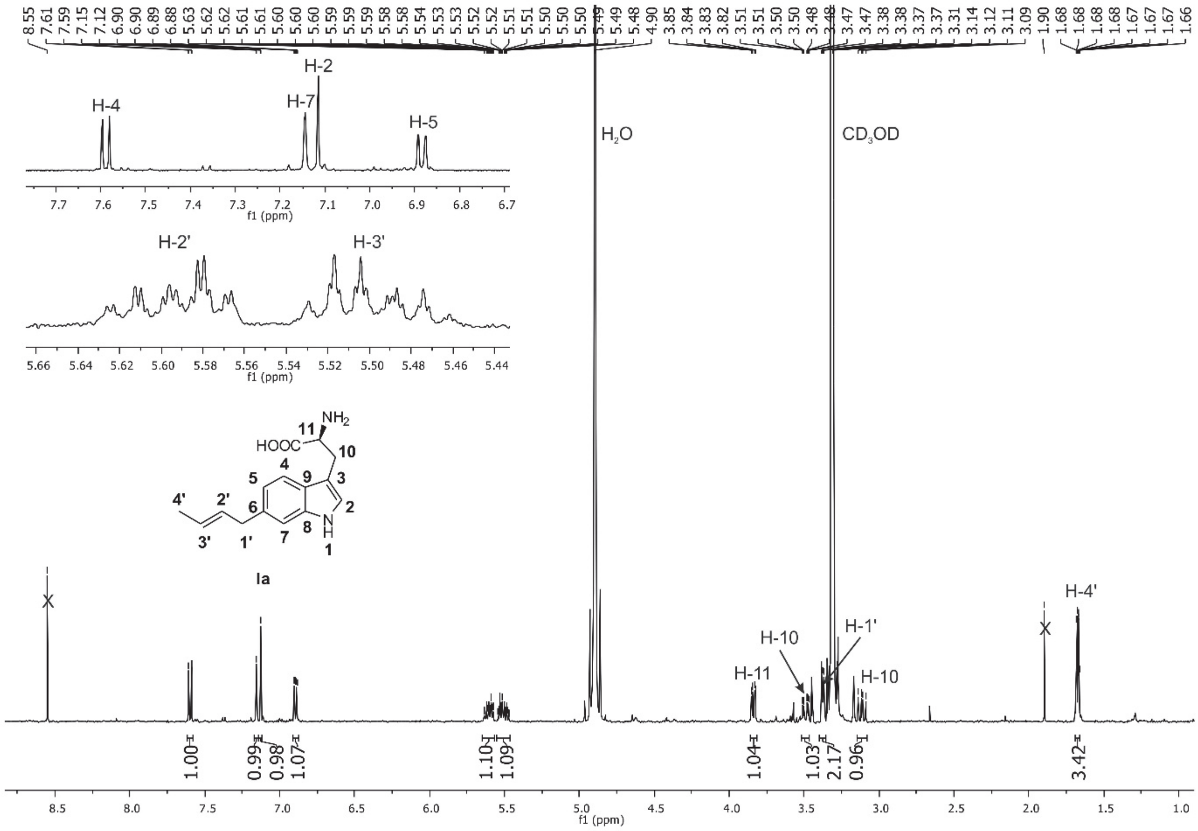


Figure S9: ^{1}H-NMR spectrum of 6-methylallyl-L-tryptophan (**Ia**) in CD_3OD.

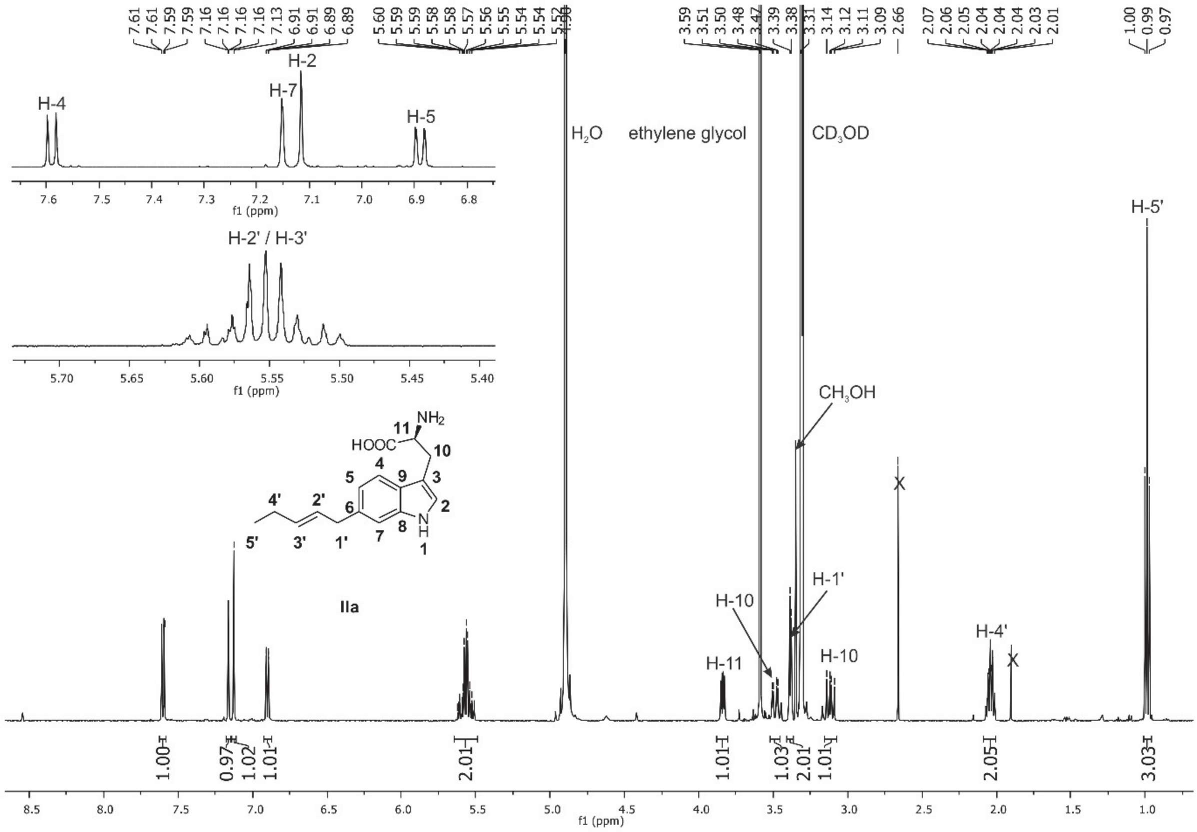


Figure S10: ^{1}H-NMR spectrum of 6-(2-pentenyl-)-L-tryptophan (**IIa**) in CD_3OD.

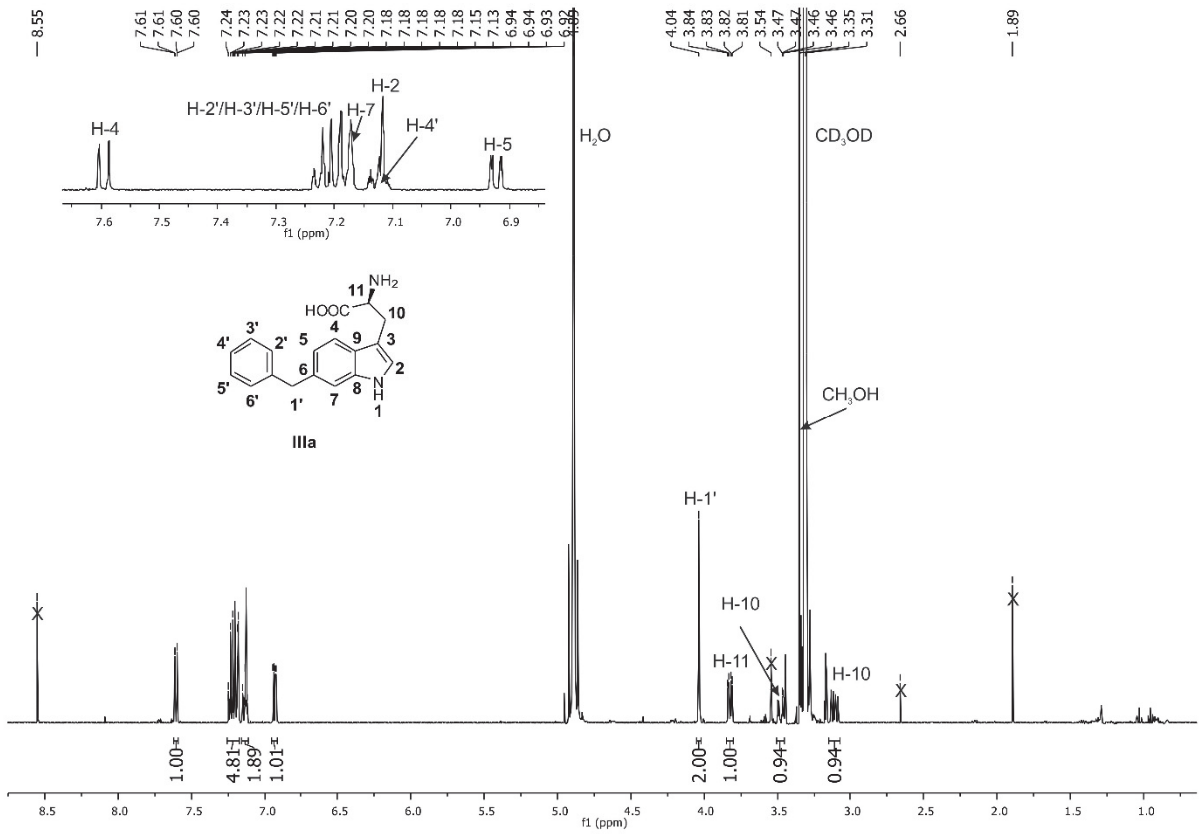


Figure S11: ^{1}H-NMR spectrum of 6-benzyl-L-tryptophan (**IIIa**) in CD_3OD.

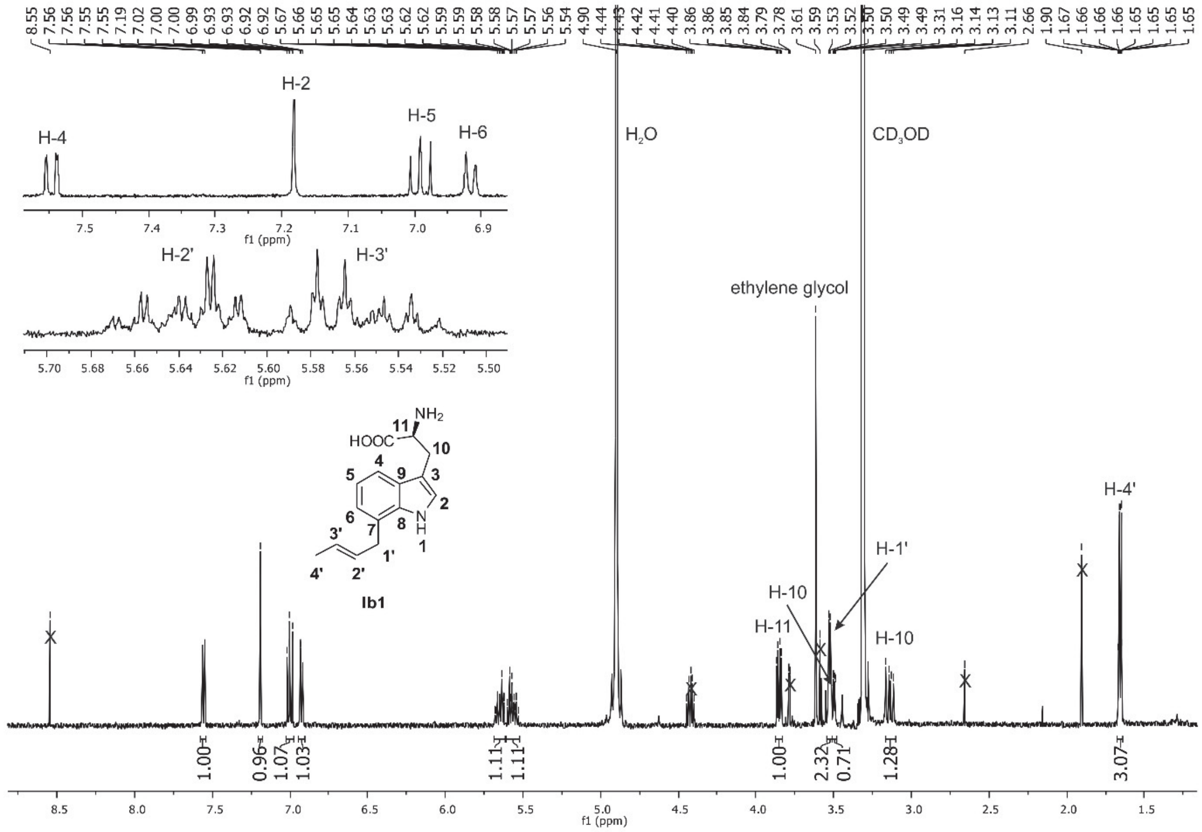


Figure S12: ^{1}H-NMR spectrum of 7-methylallyl-L-tryptophan (**Ib1**) in CD_3OD.

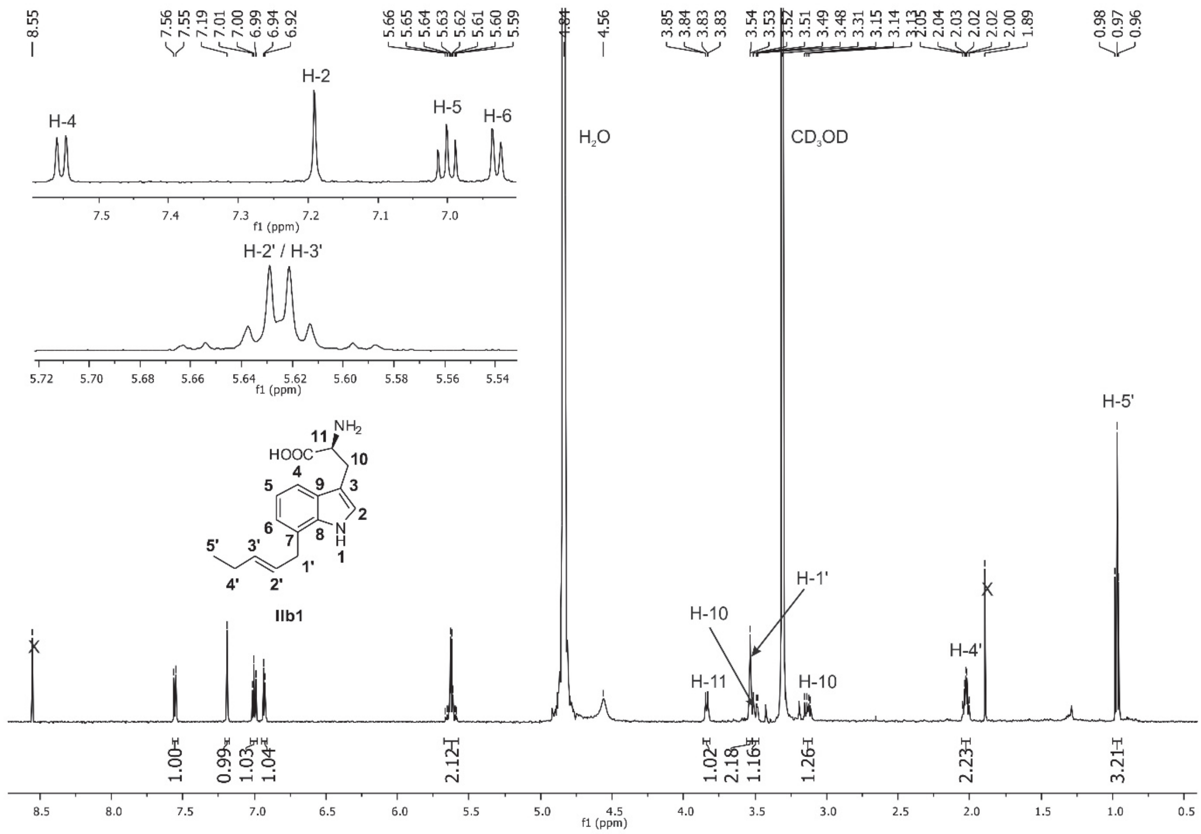


Figure S13: ^{1}H-NMR spectrum of 7-(2-pentenyl-)-L-tryptophan (**IIb1**) in CD_3OD.

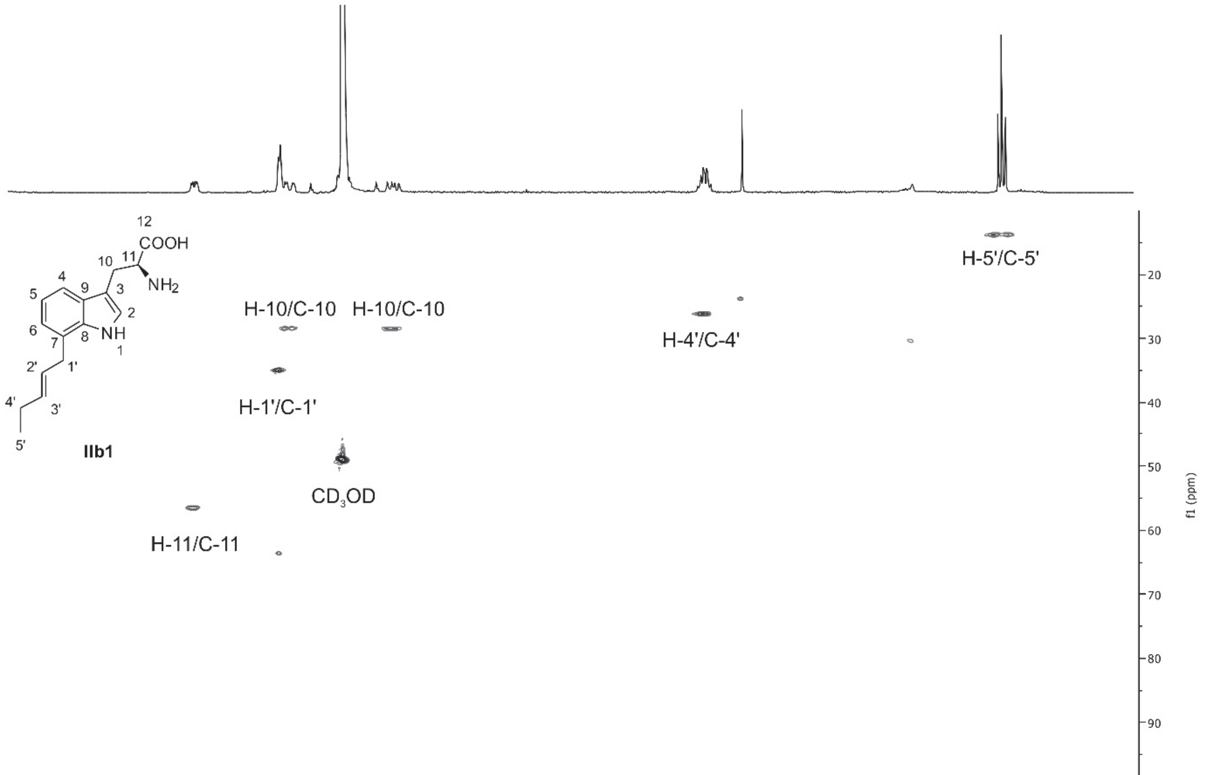


Figure S14: HSQC spectrum of 7-(2-pentenyl-)-L-tryptophan (**IIb1**) in CD_3OD.

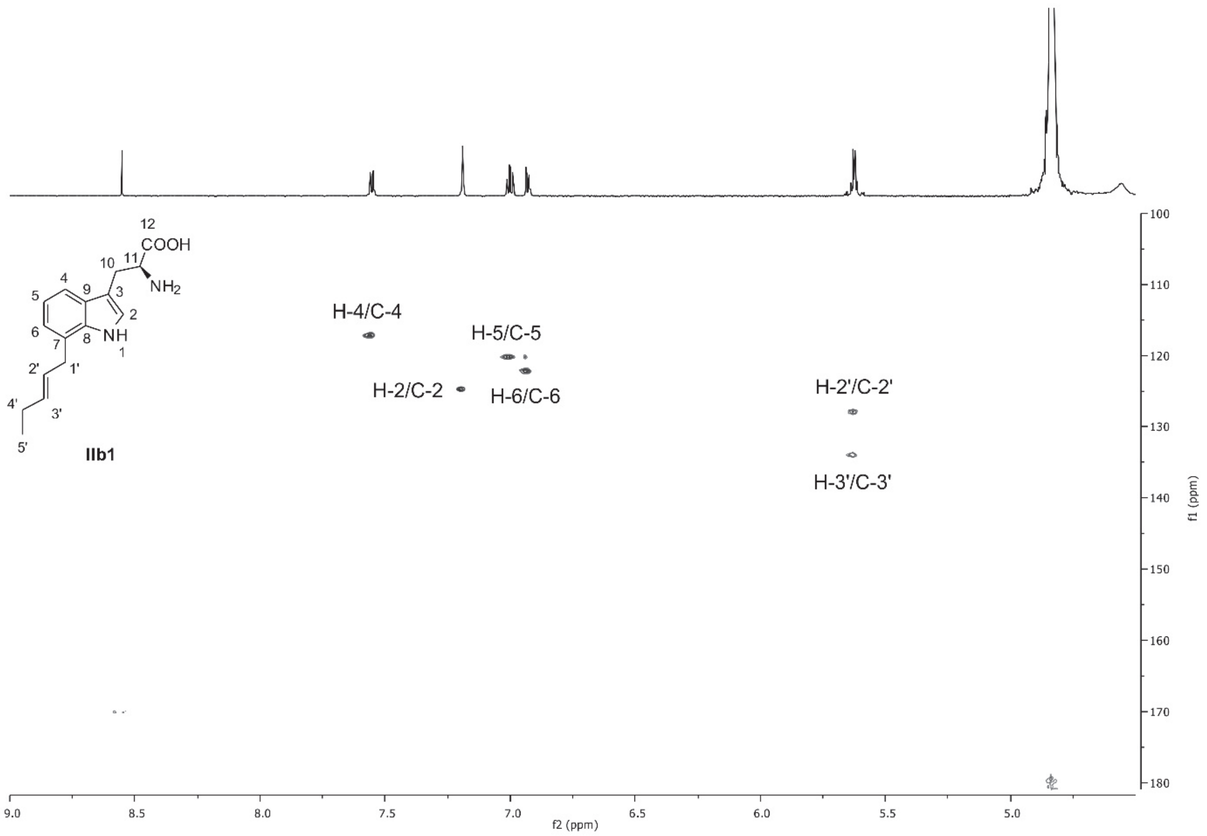


Figure S15: HSQC spectrum of 7-(2-pentenyl-)-L-tryptophan (**IIb1**) in CD_3OD.

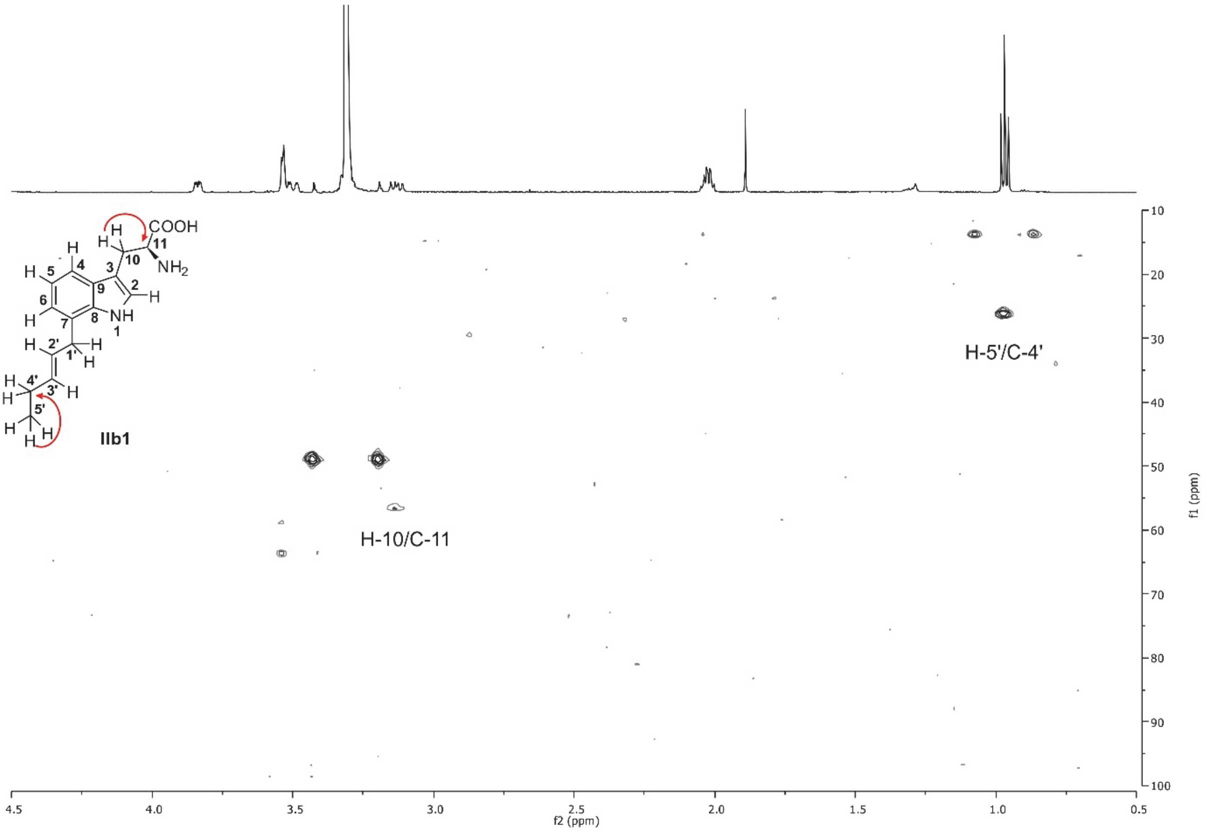


Figure S16: HMBC spectrum of 7-(2-pentenyl-)-L-tryptophan (**IIb1**) in CD_3OD.

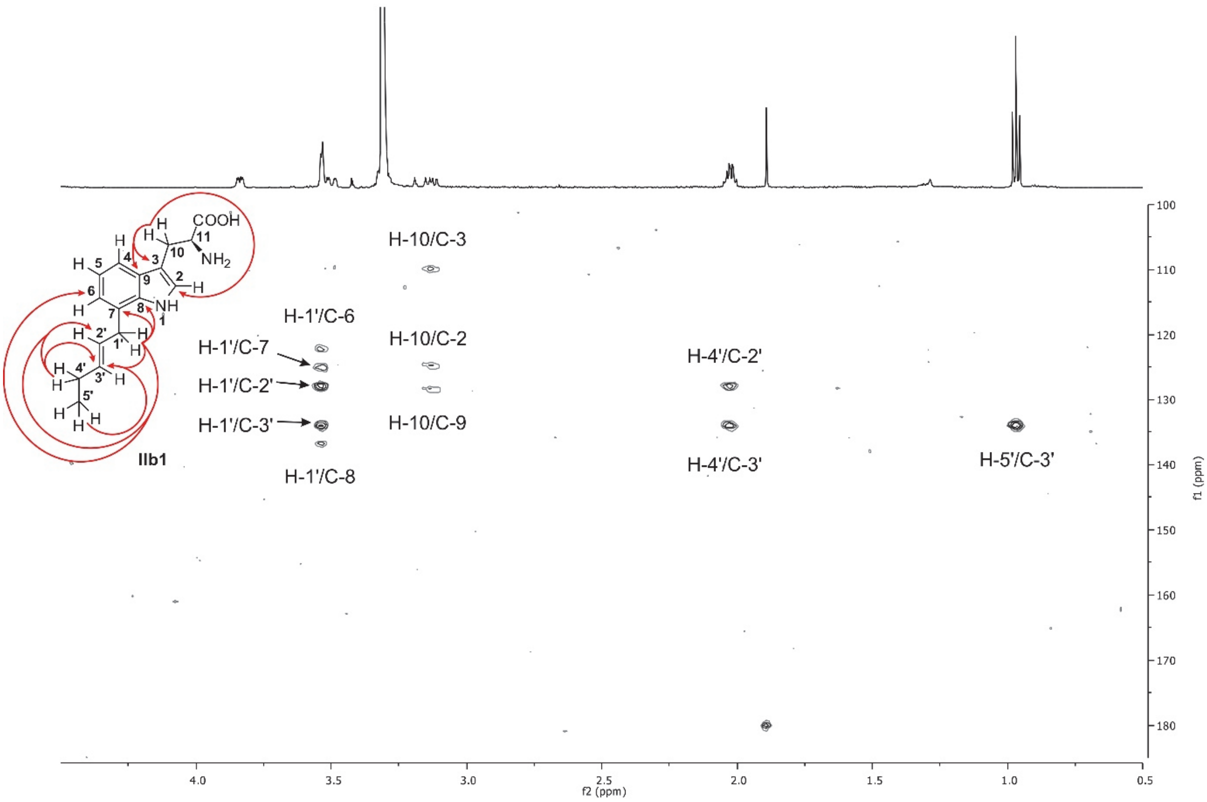


Figure S17: HMBC spectrum of 7-(2-pentenyl-)-L-tryptophan (**IIb1**) in CD_3OD.

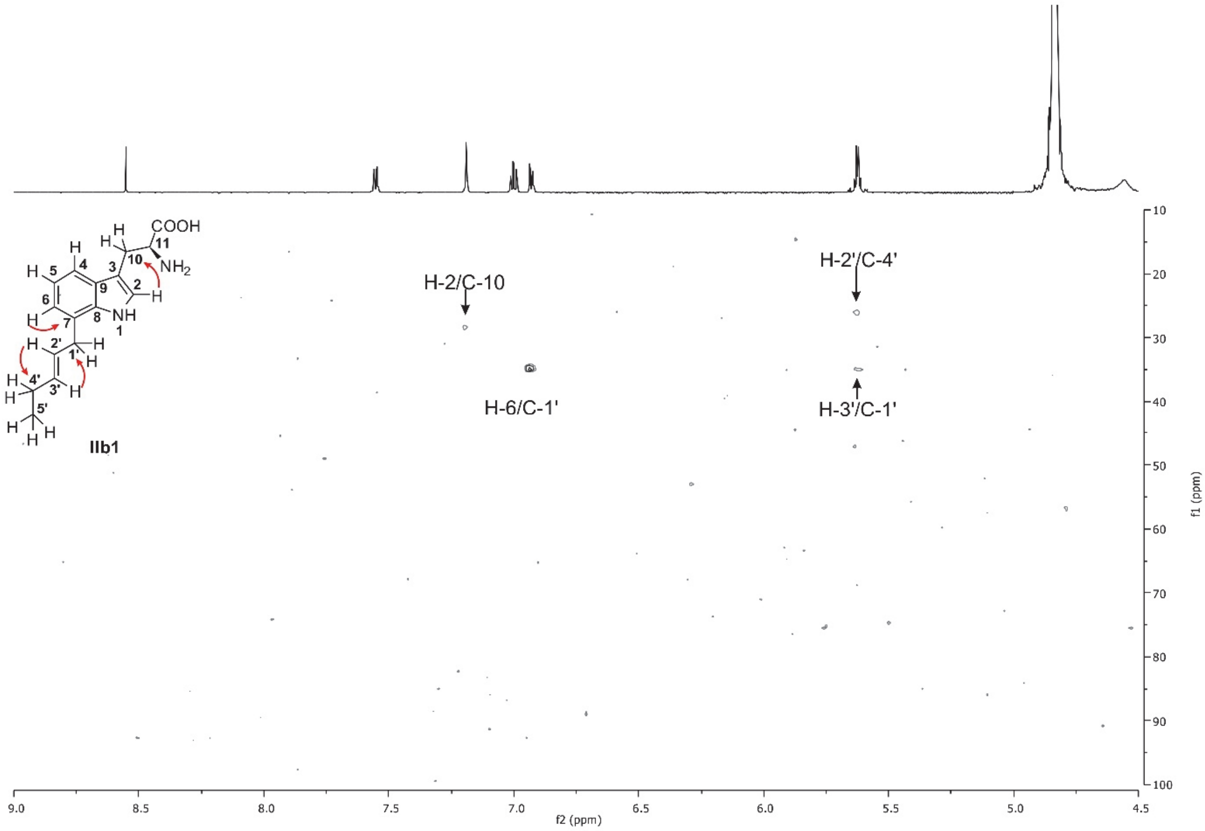


Figure S18: HMBC spectrum of 7-(2-pentenyl-)-L-tryptophan (**IIb1**) in CD_3OD.

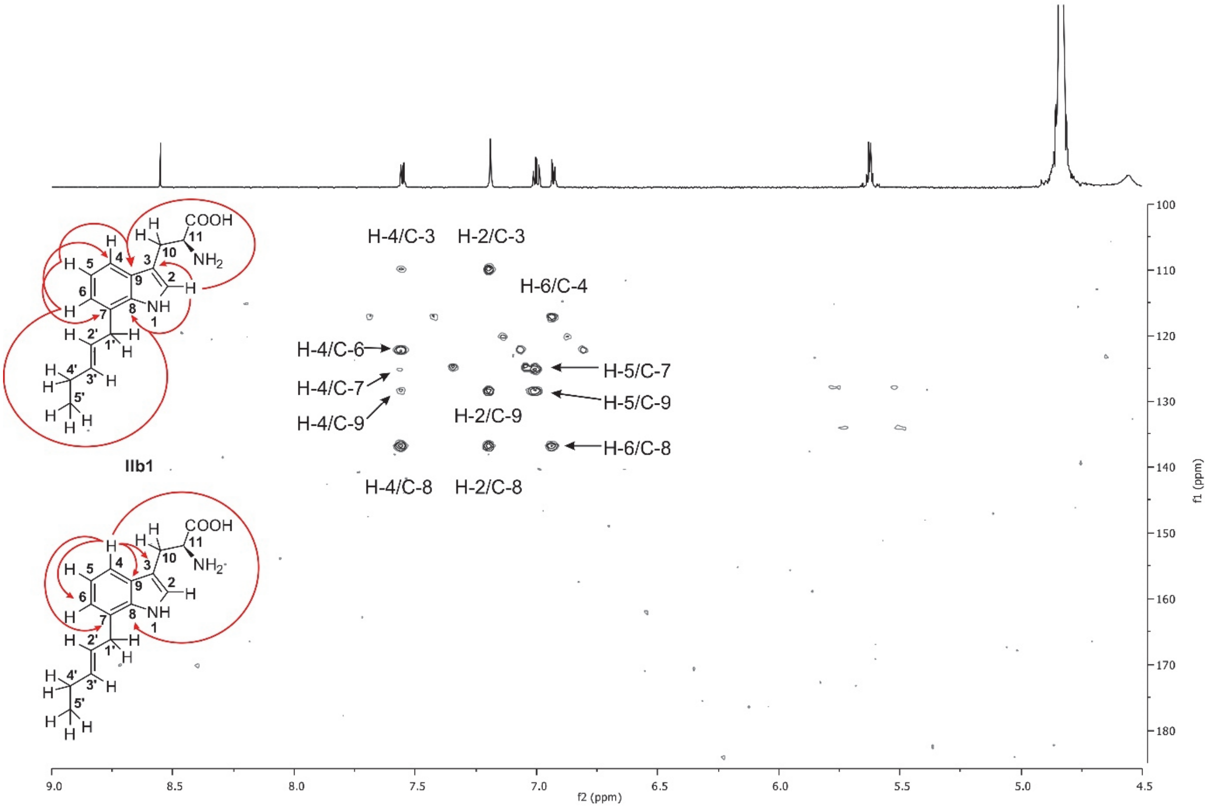


Figure S19: HMBC spectrum of 7-(2-pentenyl-)-L-tryptophan (**IIb1**) in CD$_3$OD.

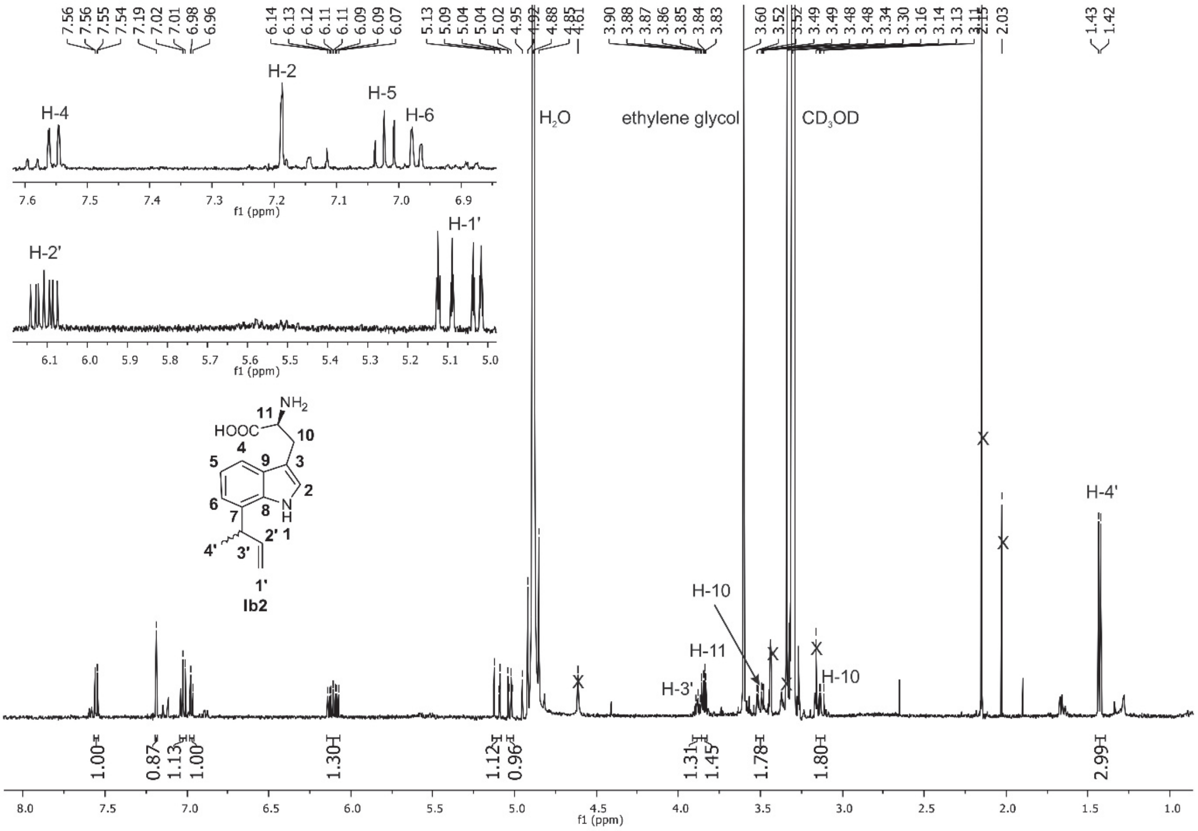


Figure S20: ^{1}H-NMR spectrum of 7-(3’*S* or 3’*R*-but-1-enyl)-L-tryptophan (**Ib2**) in CD$_3$OD.

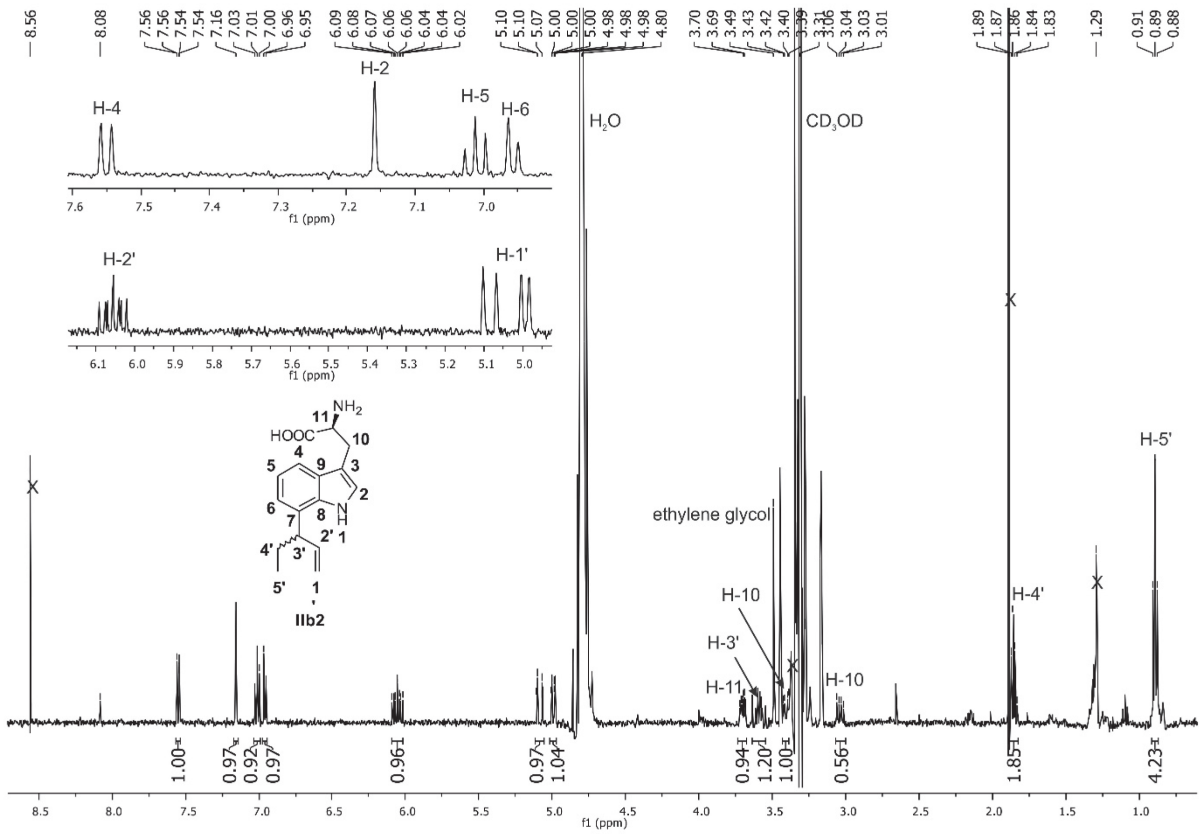


Figure S21: ^{1}H-NMR spectrum of 7-(3’*S* or 3’*R*-pent-1-enyl -)-L-tryptophan (**IIb2**) in CD_3OD.

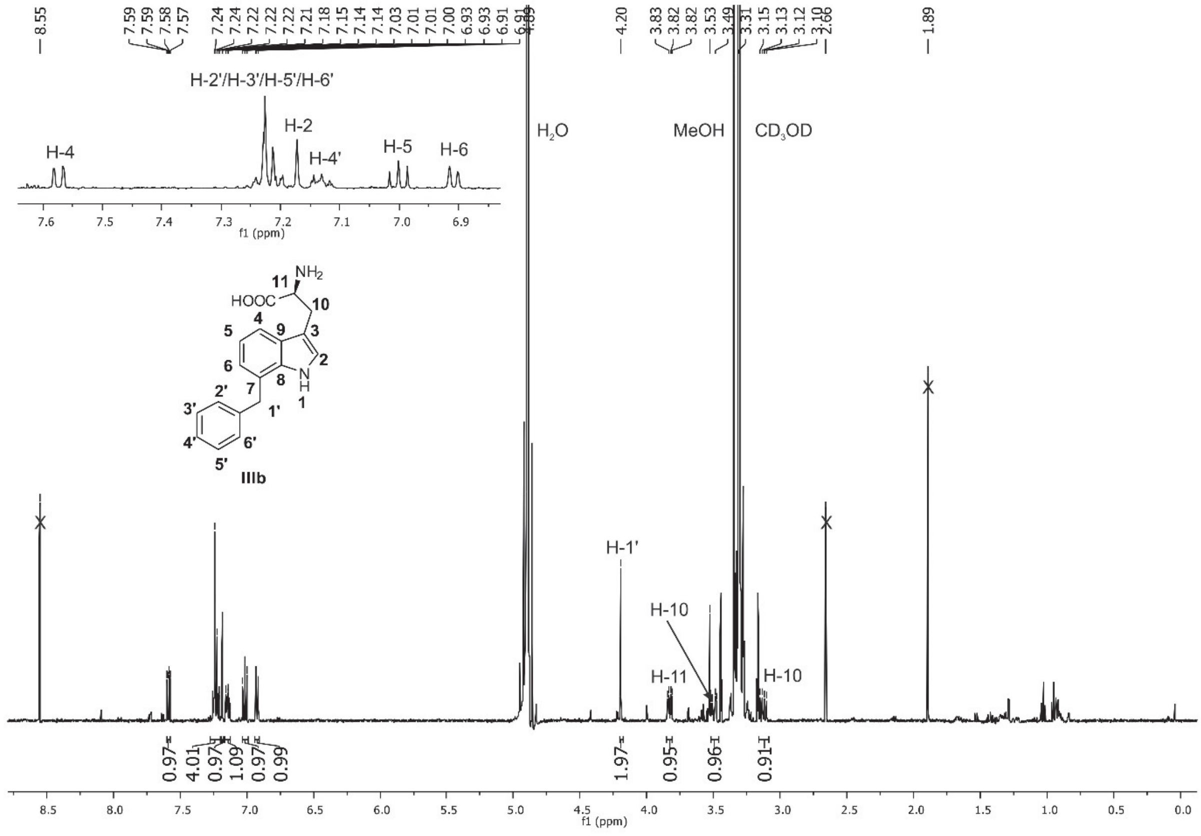


Figure S22: ^{1}H-NMR spectrum of 7-benzyl-L-tryptophan (**IIIb**) in CD_3OD.

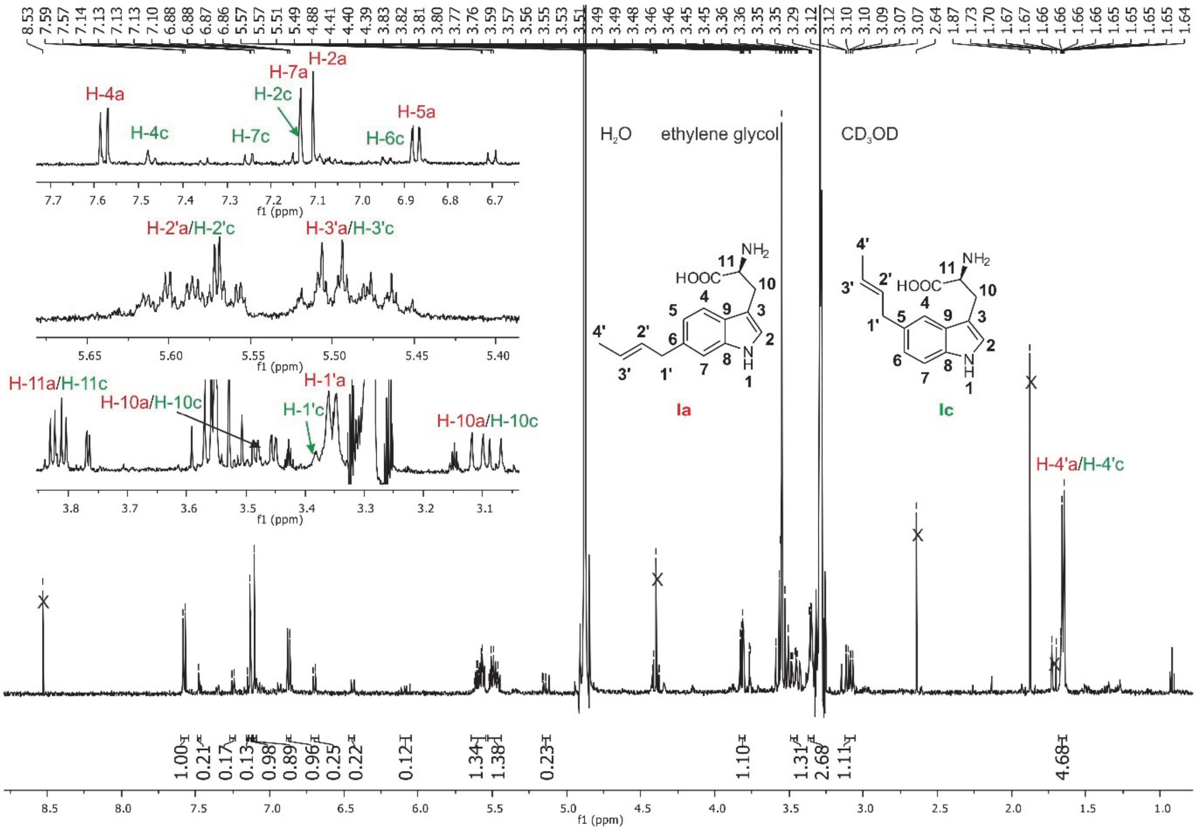


Figure S23: ^{1}H-NMR spectrum of 6-methylallyl-L-tryptophan (**Ia**) and 5-methylallyl-L-tryptophan (**Ic**) in CD$_3$OD.

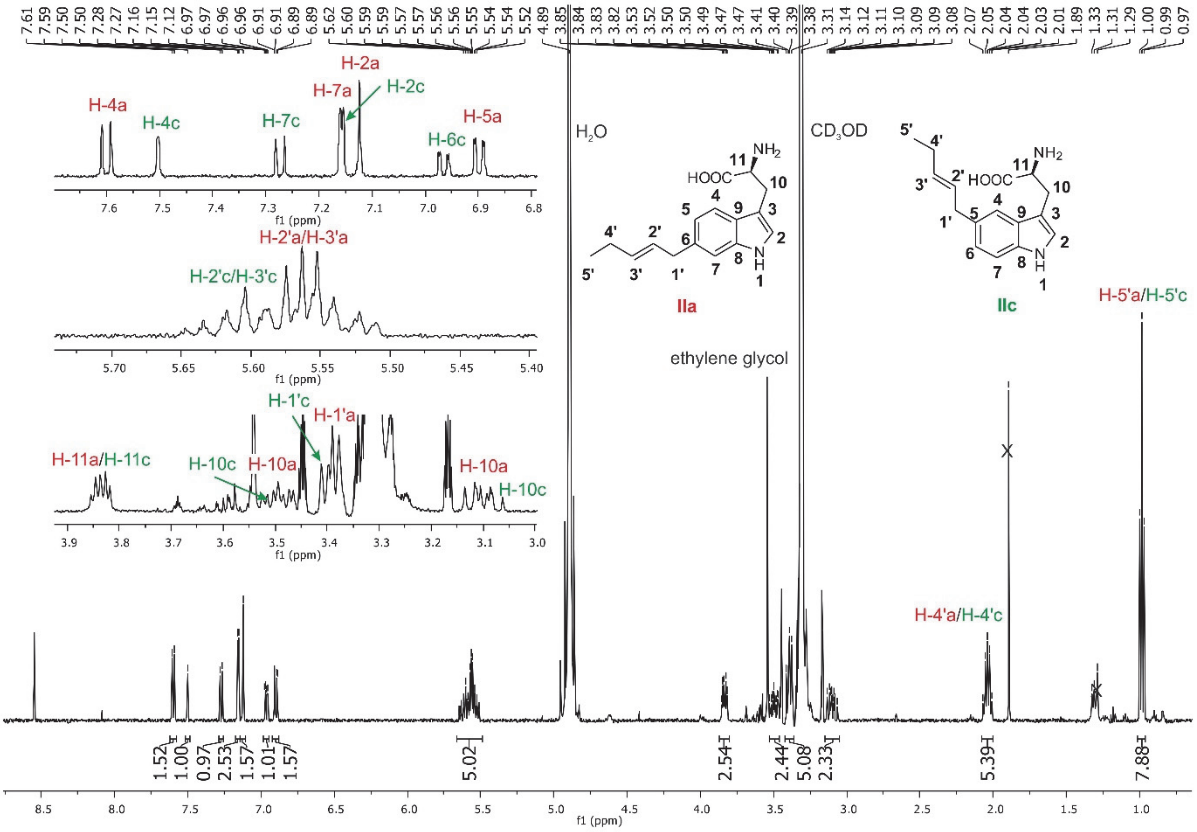


Figure S24: ^{1}H-NMR spectrum of 6-(2-pentenyl-)-L-tryptophan (**IIa**) and 5-(2-pentenyl-)-L-tryptophan (**IIc**) in CD$_3$OD.

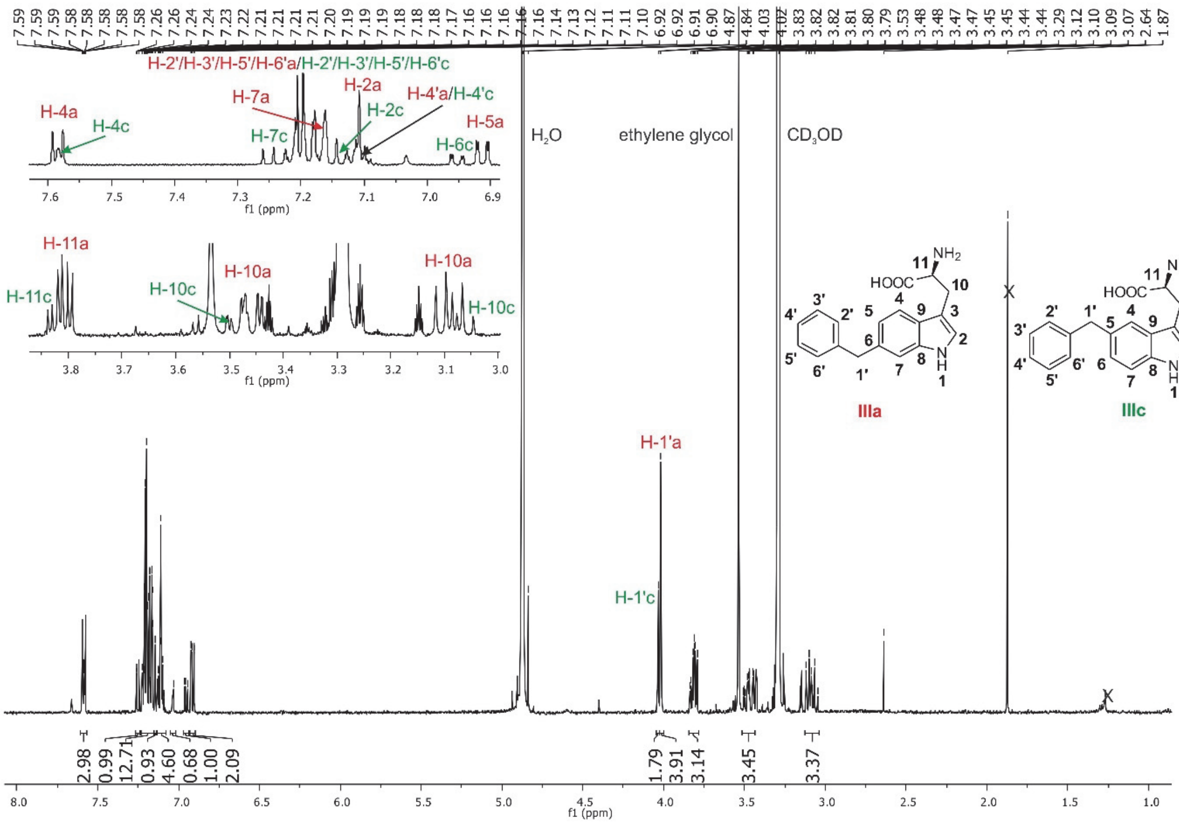


Figure S25: ^{1}H-NMR spectrum of 6-benzyl-L-tryptophan (**IIIa**) and 5-benzyl-L-tryptophan (**IIIc**) in CD$_3$OD.

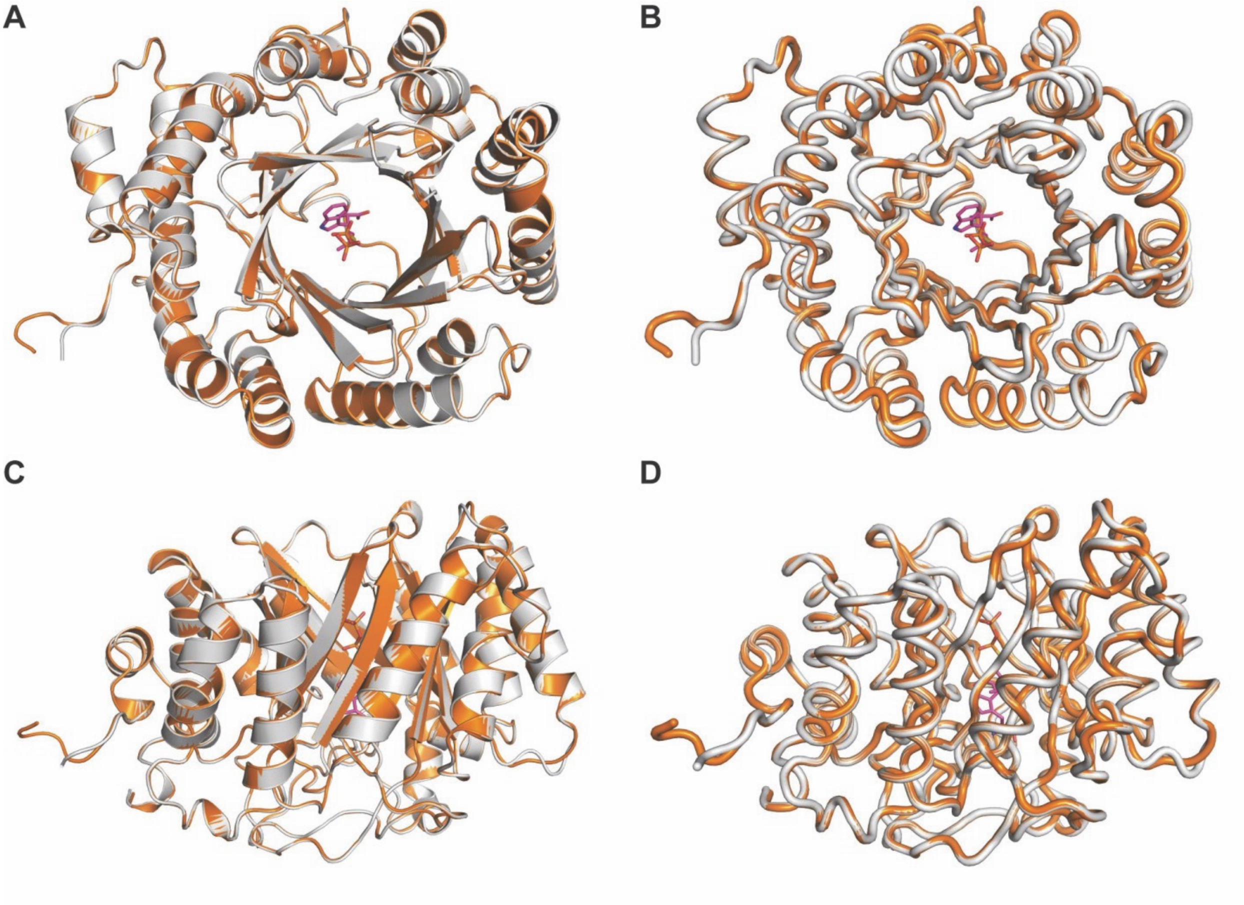


Figure S26: Homology model of 5-DMATS (orange) superimposed on the X-ray structure of FgaPT2 (white). L-tryptophan and DMAPP are shown in magenta. **A** and **C**: Cartoon representations, **B** and **D**: Cα-trace representations of template and model, respectively.

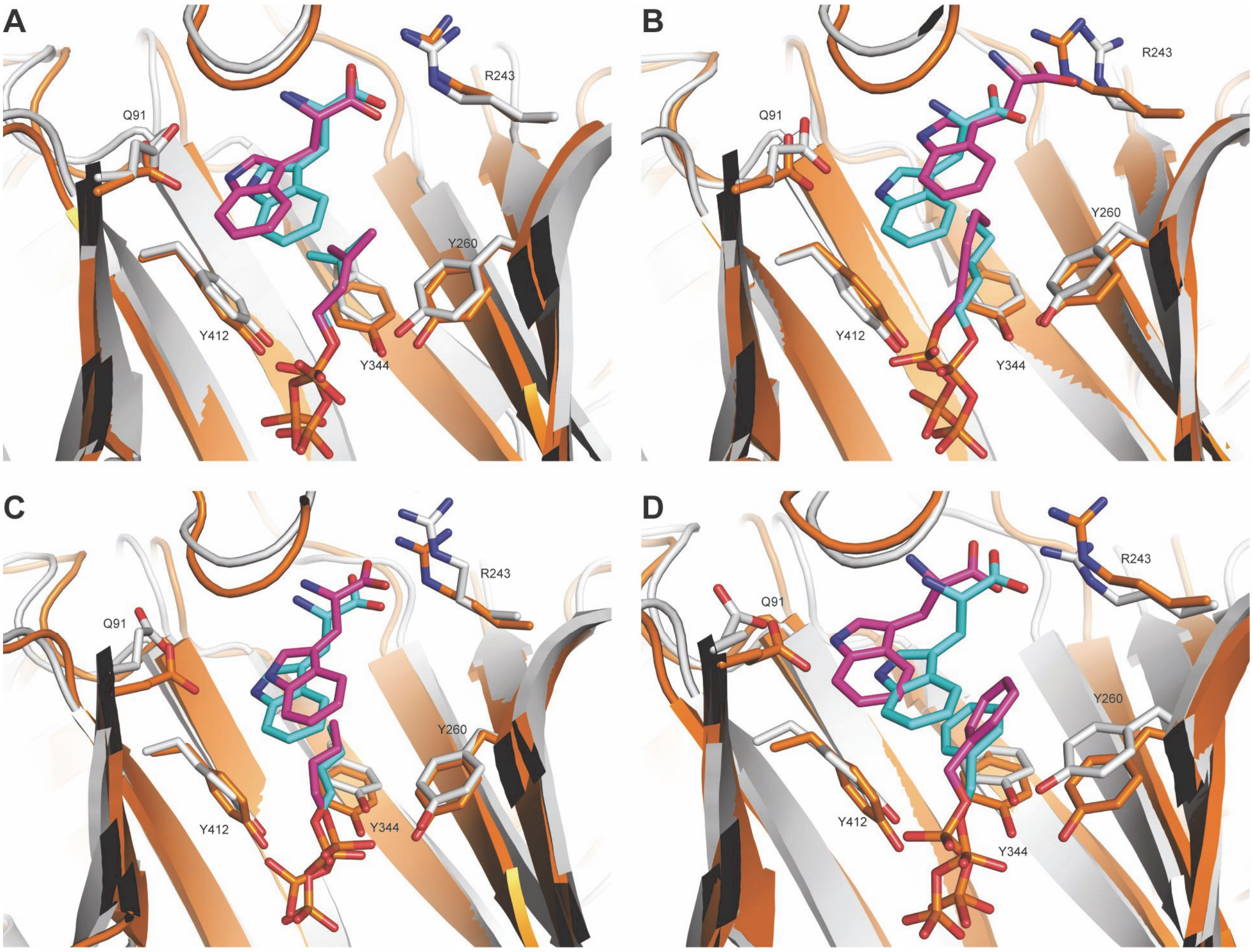


Figure S27: Shown are complexes of 5-DMATS, L-tryptophan and **A**: DMAPP, **B**: 2-pentenyl-PP, **C**: MAPP and **D**: benzyl-PP, respectively. Colored in orange and cyan is the equilibrated complex before the productive molecular dynamics run. Colored in white and magenta are complexes after 5 ns productive simulation run.

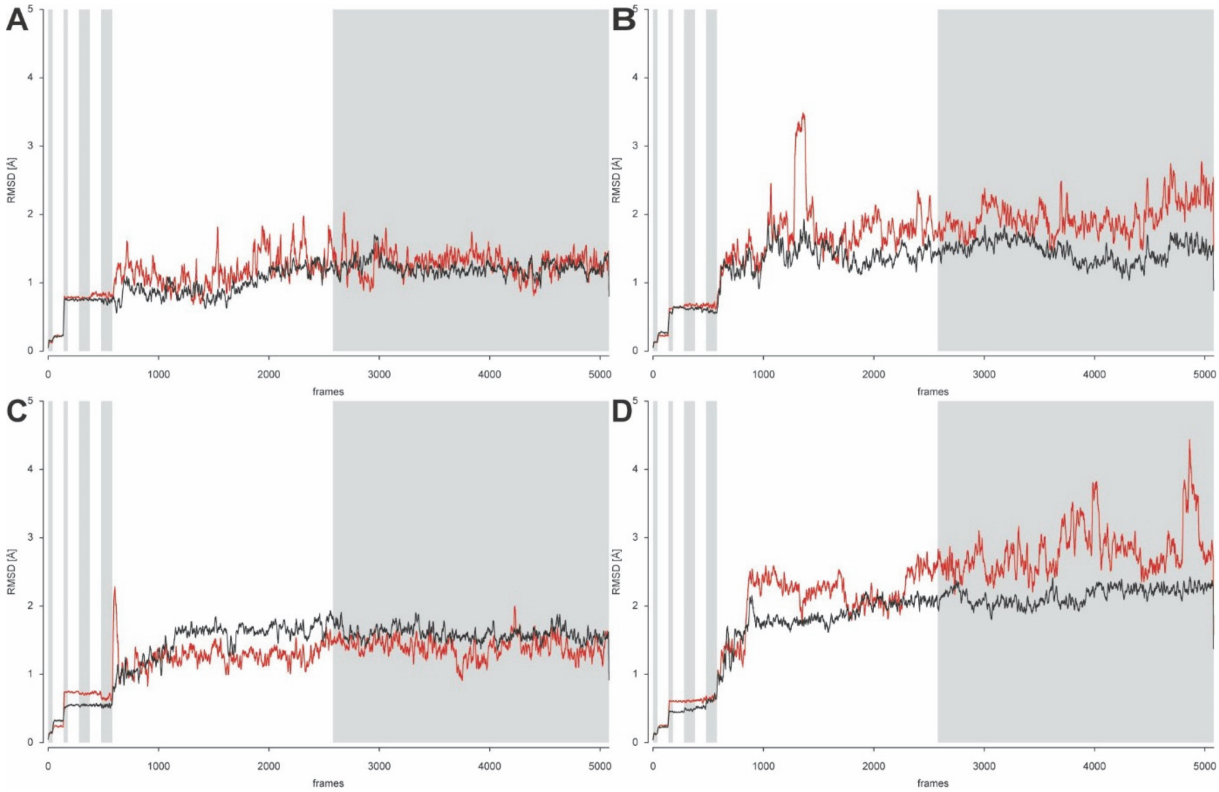


Figure S28: Shown are running averages of the RMSD of pairs of L-tryptophan (red) and a donor molecule (black) during each step of the simulations indicated by white and grey background, respectively. Tryptophan with **A**: DMAPP, **B**: 2-pentenyl-PP, **C**: MAPP and **D**: benzyl-PP.

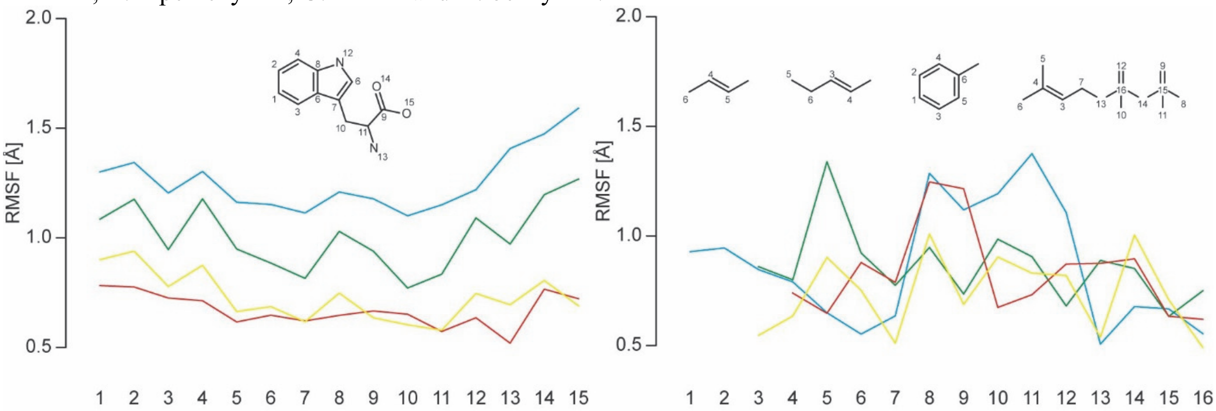


Figure S29: Shown are root-mean-square-fluctuations RMSF of L-tryptophan (TRP) (left) and donor molecules (right) during the complete simulation runs. Color-code: blue: TRP and benzyl-PP complex, green: TRP and 2-pentenyl-PP complex, yellow: TRP and DMAPP complex and red: TRP and MAPP complex.

Similarity- and substructure-based development of β2-adrenergic receptor ligands based on unusual scaffolds

Denis Schmidt[†,‡], Jakub Gunera[†], Jillian G. Baker[¶] and Peter Kolb[†,*]

[†]Department of Pharmaceutical Chemistry, Philipps-University, Marburg, Hesse, 35032, Germany

[‡]Institute of Pharmaceutical and Medicinal Chemistry, Heinrich-Heine-University, Düsseldorf, 40225, Germany

[¶]Cell Signalling, School of Life Science, Queen's Medical Centre, University of Nottingham, Nottingham NG7 2UH, U.K.

1. Substructure searches

Eight substructure queries (**S1**-**S8**), which are depicted in Chart S2, were manually derived from the six original hits (Chart S1). These queries were encoded as SMARTS and run against the complete ZINC database (*1*), which comprised 8.5 million molecules at that time.

2. Similarity searches

At the same time, FCFP4 fingerprints, as implemented in Pipeline Pilot, of the six query molecules **Q1-Q6** were used to screen the same database of 8.5 million entries for molecules with a Tanimoto similarity ≥ 0.45.

3. Docking

All molecules originating from the two searches were, after removal of duplicates, docked into the orthosteric pocket of the inverse-agonist bound X-ray structure of the β_2AR (PDB 2RH1), as previously described. (*2*) Briefly, molecules were placed by DOCK, using guiding points inside the pocket that had been derived from carazolol, the inverse agonist bound to the β_2AR in this X-ray structure.

4. Cell culture

CHO-K1 cells stably expressing either the human β_1AR or the human β_2AR and CRE-SPAP reporter gene were used (CHO-β_1, CHO-β_2 (*3*)) and grown in Dulbecco's modified Eagle's medium nutrient mix F12 (DMEM/F12) containing 10% foetal calf serum and 2 mM l-glutamine in a 37°C humidified 5% CO_2:95% air atmosphere.

5. [^{3}H](-)CGP 12177 whole cell binding

Media was removed from confluent cells in white-sided 96-well plated and immediately replaced by 100 µL of the ligand under investigation (diluted in serum-free media (sfm), DMEM/F12 containing 2 mM l-glutamine only) followed immediately by 100 µL [^{3}H](-)CGP 12177 (in sfm) to give a final [^{3}H](-)CGP 12177 concentration of 0.44-1.12 nM. The plates were the incubated for 2 h at 37°C before being washed twice with 200 µL 4°C phosphate buffered saline. Microscint 20 (100 µL) was added to each well, a white base added to the plate, the plated left for a minimum of 8 h in the dark then counted on a TopCount.

K_D values were determined from the IC_{50} values using the Cheng-Prusoff equation (see ref. (*3*) for details). For all ligands that completely inhibited specific binding, a pK_D value is given. For ligands where significant specific binding was inhibited, but the maximum concentration of a ligand was not quite sufficient to completely inhibit specific binding, an apparent pK_D value is given (based on the assumption that a higher concentration of the competing ligand would inhibit all specific binding). For ligands with less than 50% inhibition of binding, despite maximum concentration of ligand (maximum possible concentration of ligand ranged from 20-100 µM), no K_D value is stated. Propranolol (10 µM) was used to determine non-specific binding and the K_D values for $[^3H]$(-)CGP 12177 were 0.42 nM for the human β_1AR and 0.17 nM for the human β_2AR. (*3*)

6. CRE-SPAP production

Confluent cells (96-well plates) were serum starved with sfm for 24 h before experimentation. The media was then removed and replaced with 100 µL sfm or sfm containing final concentration of antagonist. Agonist (10 µL, diluted in sfm) was then added and the plates incubated for 5 h at 37°C.

CRE-SPAP production was then measured as previously described. (*4*)The intrinsic efficacy of all ligands was assessed from 7-point concentration response curves. Isoprenaline (10 µM) was used as the positive control in all plates. Maximum responses and pEC_{50} values were obtained from sigmoidal dose response curves (see ref. (*4*) for full details). The affinity of antagonists was determined from a rightward shift of the agonist response using the Gaddam equation, and for the partial agonist **3**, using the method of Stephenson (see ref. (*4*) or full details).

7. Supplementary Results

7.1. $[^3H]$(-)CGP 12177 whole cell binding and CRE-SPAP production validation

$[^3H]$(-)CGP 12177 whole cell binding demonstrated that the known β_1-selective antagonist CGP 20712A, as expected, had high affinity for the human β_1AR (pK_D 8.96±0.13, n=4) whilst the known β_2-antagonist ICI 118551 had high affinity for the human β_2AR (pK_D 9.61±0.05, n=5, Table 1).

Cimaterol stimulated a full agonist response at both receptors. At the β_1AR, this response was 3.3±0.5-fold over basal, 105±2% that of the isoprenaline maximum (n=12) and at the β_2AR, the response was 4.4±0.1-fold over basal and 95±1% that of isoprenaline (n=9) (Table S3). As expected, CGP 20712A inhibited the CHO-β_1 cimaterol response with high affinity, and ICI 118551 inhibited the CHO-β_2 cimaterol response with high affinity to yield similar selectivities to those obtained from the binding assay (Table 2).

7.2. Compound 3 acts through the primary catecholamine conformation of β_1AR

Compound **3** was clearly a partial agonist at both the β_1 and β_2-AR (Table S4, Figure S2). At the β_2AR, this partial agonist response was inhibited by ICI 118551 to give a K_D value for ICI 118551 very similar to that obtained in the presence of cimaterol (Table S4), confirming that this partial agonist response is indeed occurring through interaction with the β_2AR.

The β_1AR, however, exists in at least two active agonist conformations: (*4-8*) a high affinity catecholamine conformation (through which cimaterol and catecholamines stimulate agonist responses, and for which CGP 20712A and CGP 12177 have high affinities), and a secondary conformation through which higher concentrations of CGP 12177 stimulate agonist responses (although these responses are relatively resistant to antagonism). The conformation through which **3** was stimulating β_1 partial agonist response was therefore assessed.

The affinity measured by both the binding assay (pK_D 9.01) and the functional assay (pK_D 9.19) were very similar. The concentrations of $[^3H]$(-)CGP 12177 used in the binding assay would only measure binding to the catecholamine conformation. Compound **3** also inhibited the cimaterol response (known to act through the catecholamine conformation, (*4-7*) with high affinity, again suggesting high affinity catecholamine conformation interaction. The partial agonist response (pEC_{50} 8.80) is also very similar to the pK_D values, again suggesting interactions with the catecholamine conformation. This therefore suggests that both the binding of **3** and the agonist response obtained in the functional assay are occurring through the same high affinity conformation of the β1AR.

The partial agonist response of **3** in CHO-β1 cells was inhibited by CGP 20712A with high affinity, suggesting that the response is indeed β_1AR-mediated. However, the K_D value for CGP 20712A obtained was part-way between that of cimaterol (high affinity catecholamine conformation) and that of CGP 12177 (secondary conformation, Table S4). Thus, although the similarity of the K_D and EC_{50} values suggests single-site, high affinity conformation interactions, further evidence for which site of actions the response was occurring through was sought. When increasing concentrations of **3** were added to fixed concentrations of cimaterol (Figure S2), the cimaterol response was inhibited in a manner suggestive of competition at a single conformation (compare with Figure 1 of ref (*5*); Figure 4 of ref (*6*); Figure 8 of ref (*7*) and Figure 6 of ref (*4*)). Overall therefore, **3** is also a high affinity partial agonist of the human β_1AR, with the agonist response occurring through the primary catecholamine conformation of the receptor.

7.3. Dose response curves of the other ligands

The dose response curve for several compounds showed no stimulation of either receptor (e.g. Figure S3, compound **1**). For some ligands, e.g. **16** and **17**, there was also no inhibition of [^{3}H](-)CGP 12177 binding and no shift of the cimaterol-induced concentration response curve. These ligands were therefore found to not be interacting with either the β_1 or β_2AR at concentrations up to the maximum studies (100 μM for many). Other compounds, e.g. **1**, although no stimulation occurred in response to the ligands alone, they did inhibit binding and cause a shift of the cimaterol-induced dose response. These compounds are therefore neutral antagonists.

For some compounds, e.g. **10** and **11**, the highest concentrations possible (20 μM for **10**, 100 μM for **11**) caused a marked fall to below basal (e.g. Figure S3). This pattern of CRE-SPAP production is consistent with toxicity (i.e. cell death, or major assay interference). In these instances, the concentration of compound used to antagonize cimaterol was reduced, until such a time as the reduction in basal was minimal or non-existent (i.e. for **10**, reduced to 2 μM as this no longer caused a reduction in basal). The functional assay is far more sensitive to issues such as toxicity because the cells need to be living in order to generate responses, whereas in the binding assay, binding to the receptors can be measured even if the cells are dead. In some cases, this reduction in compound concentration still allowed a shift to be observed and thus a K_D value to be measured.

Receptor-mediated inverse agonism as an explanation for the marked fall in CRE-SPAP production is very unlikely as i) this gene transcription assay is relatively poor at detecting inverse agonism, including compound ICI 118551, which is known to be an inverse agonist in these cells; (*9*) ii) identical results were seen in both β_1- and β_2-cells despite the fact that the ligands e.g. **10** had different affinities for the two receptors (and therefore receptor mediated effects should have been observed at different doses); iii) the $logIC_{50}$ of the apparent fall in CRE-SPAP production (e.g. for **10**, 10μM at the β_2AR) is not the same as the K_D value obtained from the binding studies (1μM), again suggesting the fall is a non-receptor mediated issue and iv) if the fall below basal was due to inverse agonism, there should still be a cimaterol concentration response in the presence of 20μM **10**, that was further right shifted, than that at 2 μM (Figure S1). As can be seen in Figure S4, there is absolutely no cimaterol response in the presence of 20 μM **10** and the whole response is below basal. This strongly suggests a non-receptor mediated cause for the fall.

7.4. Novel compounds

As well as a small fall in maximum cimaterol response, **10** causes a rightward shift in the cimaterol concentration response at the β_2AR but not at the β_1AR. This suggests that **10** is indeed interacting with the β_2AR in this functional assay and it is showing some β_2AR selectivity, with an apparent K_D value that is similar to that obtained from the binding study. The lack of a rightward shift of the cimaterol-concentration response in the presence of 2 µM **10** at β_1AR are entirely as expected, given the K_D value obtained from the binding studies (3 µM). Thus, despite the apparent toxicity issues at high concentrations in the functional assay, **10** appears to be a β_2-selective ligand with an affinity of 300-1000 nM in both the binding and functional assay.

8. Supplementary Figures

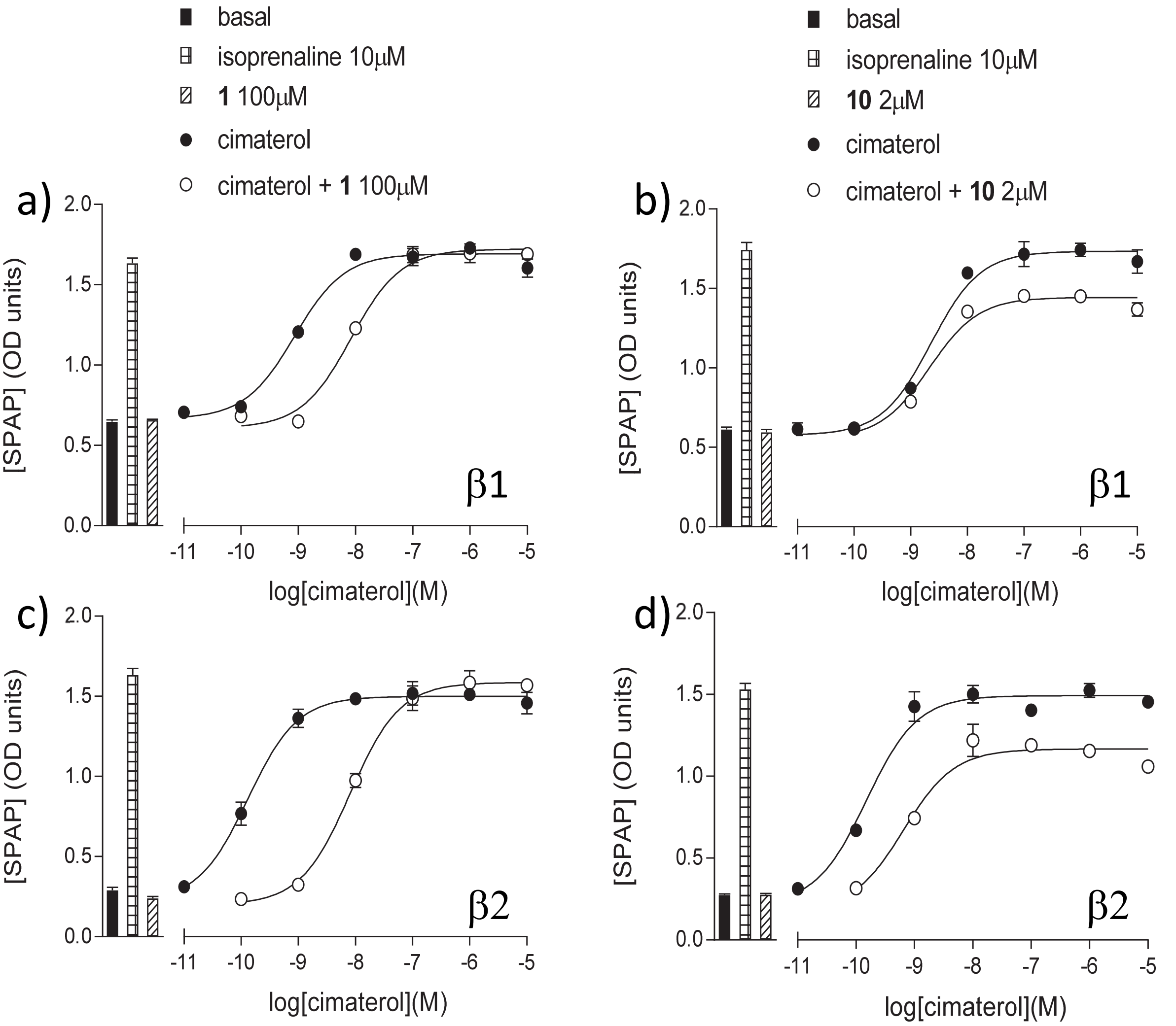


Figure S1: CRE-SPAP production in a) and b) CHO-β1 cells and c) and d) CHO-β_2 cells in response to cimaterol in the absence and presence of a) and c) 100 µM **1,** and b) and d) 2mM **10**. Bars represent basal CRE-SPAP production and that in response to 10 µM isoprenaline and 100 µM **1** or 2mM **10** alone. Data points are mean ± sem of triplicate values and these individual experiments are representative of a) and c) 6 separate experiments and b) and d) 3 separate experiments.

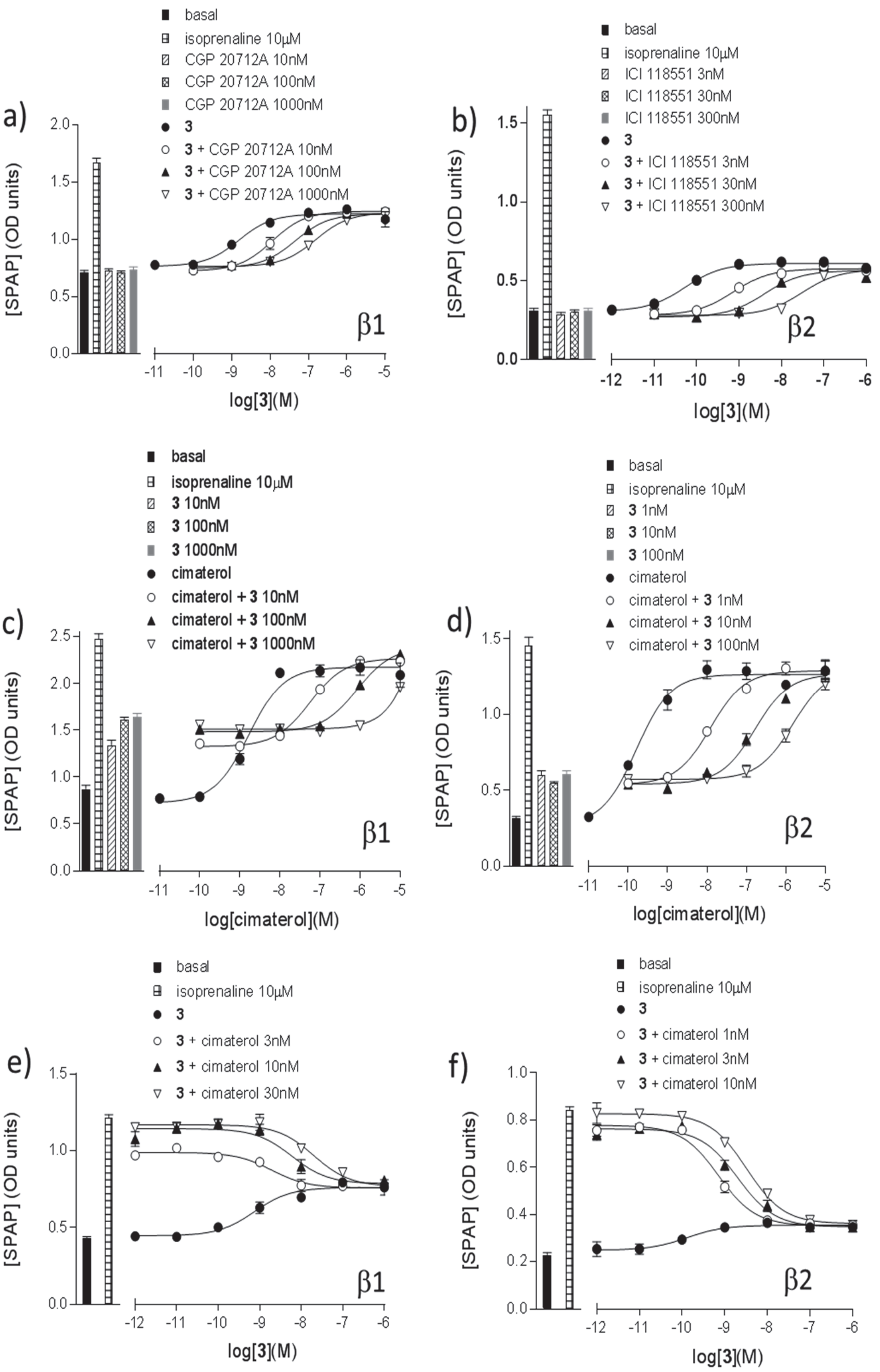


Figure S2: CRE-SPAP production in CHO-β_1 cells (a, c and e) and CHO-β_2 cells (b, d and f). a and b) show response to **3** inhibited by CGP20712A in the β_1 cells and inhibited by ICI 118551 in the β_2 cells thus confirming the responses are mediated via the respective receptors. c and d) show inhibition of the cimaterol response by increasing concentration of **3** in a manner consistent with that of a partial agonist; e and f) show inhibition of the cimaterol response by **3** in a manner consistent with competition at a single site. Bars represent basal CRE-SPAP production and that in response to 10 µM isoprenaline or various concentrations of CGP 20712A, ICI 118551 or **3** alone. Data points are mean ± sem of triplicate values and these individual experiments are representative of five or more separate experiments in each case.

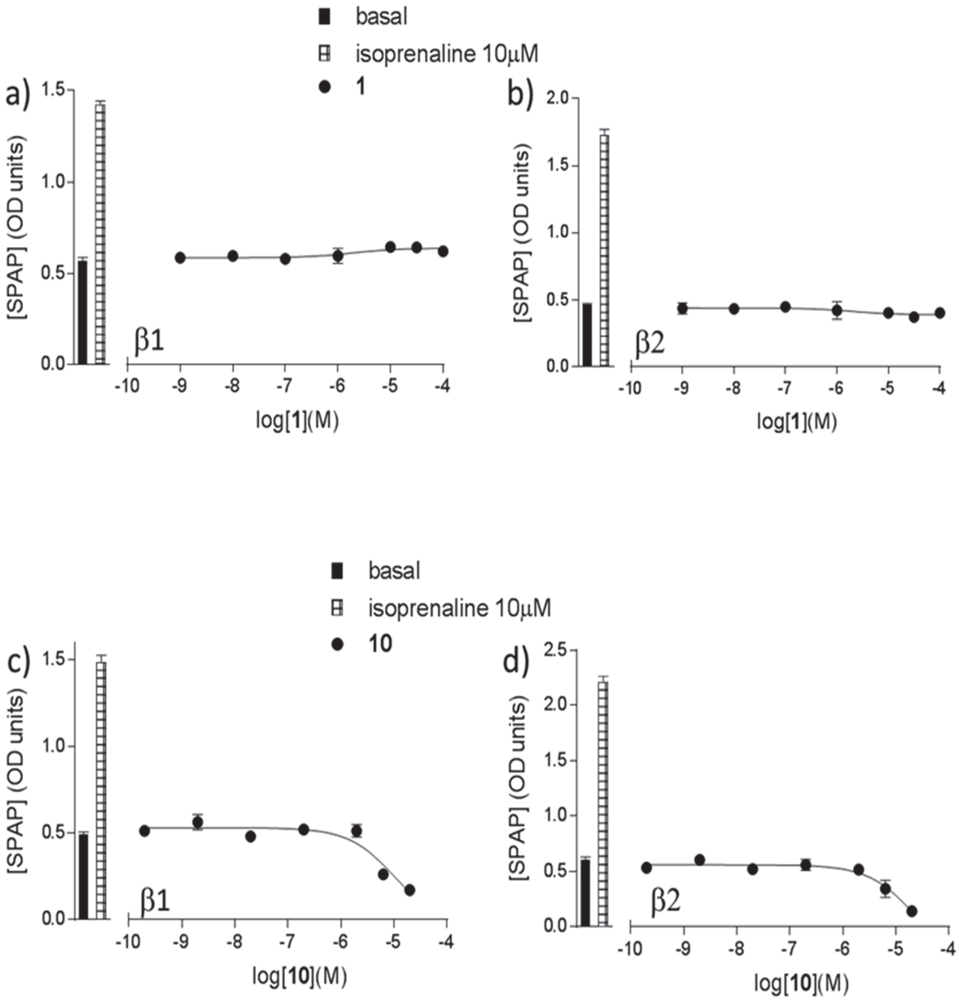


Figure S3: CRE-SPAP production in a) and c) in CHO-β_1 cells and b) and d) in CHO-β_2 cells in response to **1** (a and b) and **10** (c and d). Bars represent basal response and that to10 μM isoprenaline. Data points are mean ± sem of triplicate determinations and these individual experiments are representative of 4 separate experiments in each case.

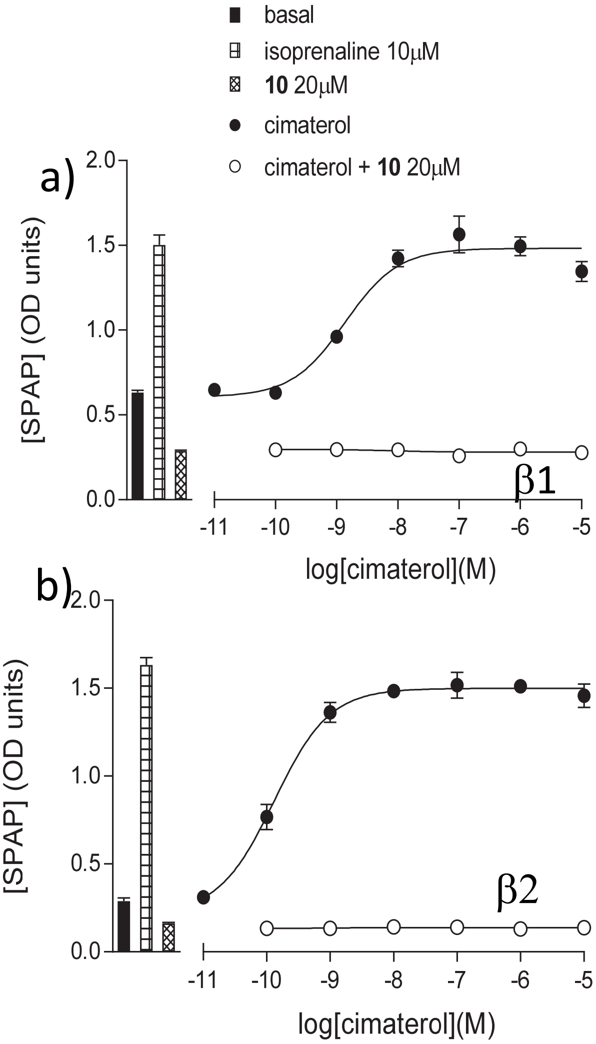


Figure S4: CRE-SPAP production in a) in CHO-β_1 cells and b) in CHO-β_2 cells in response to cimaterol in the presence and absence of 20μM **10**. Bars represent basal response and that in response to 10μM isoprenaline or 20μM **10** alone. Data points are mean ± sem of triplicate determinations and are representative of three separate experiments in each case.

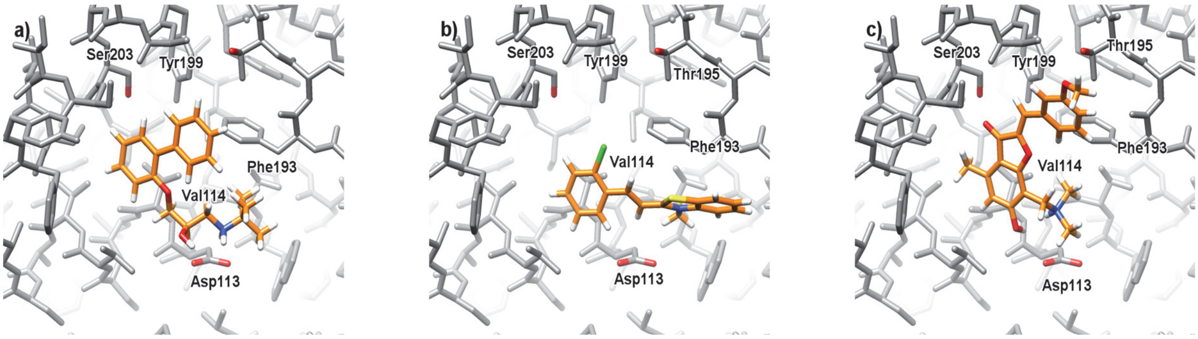


Figure S5: Docking poses for selected compounds. The β_2AR is shown in gray stick representation. Residues discussed in the text are labeled and shown with colored heteroatoms. Selected residues in TM6 and TM7 (including Phe289$^{6.51}$ and Phe290$^{6.52}$) are hidden for clarity. Ligands are shown in orange stick representation. Perspective as in ref. (5) for comparability. (a) **3**, (b) **11**, (c) **7**.

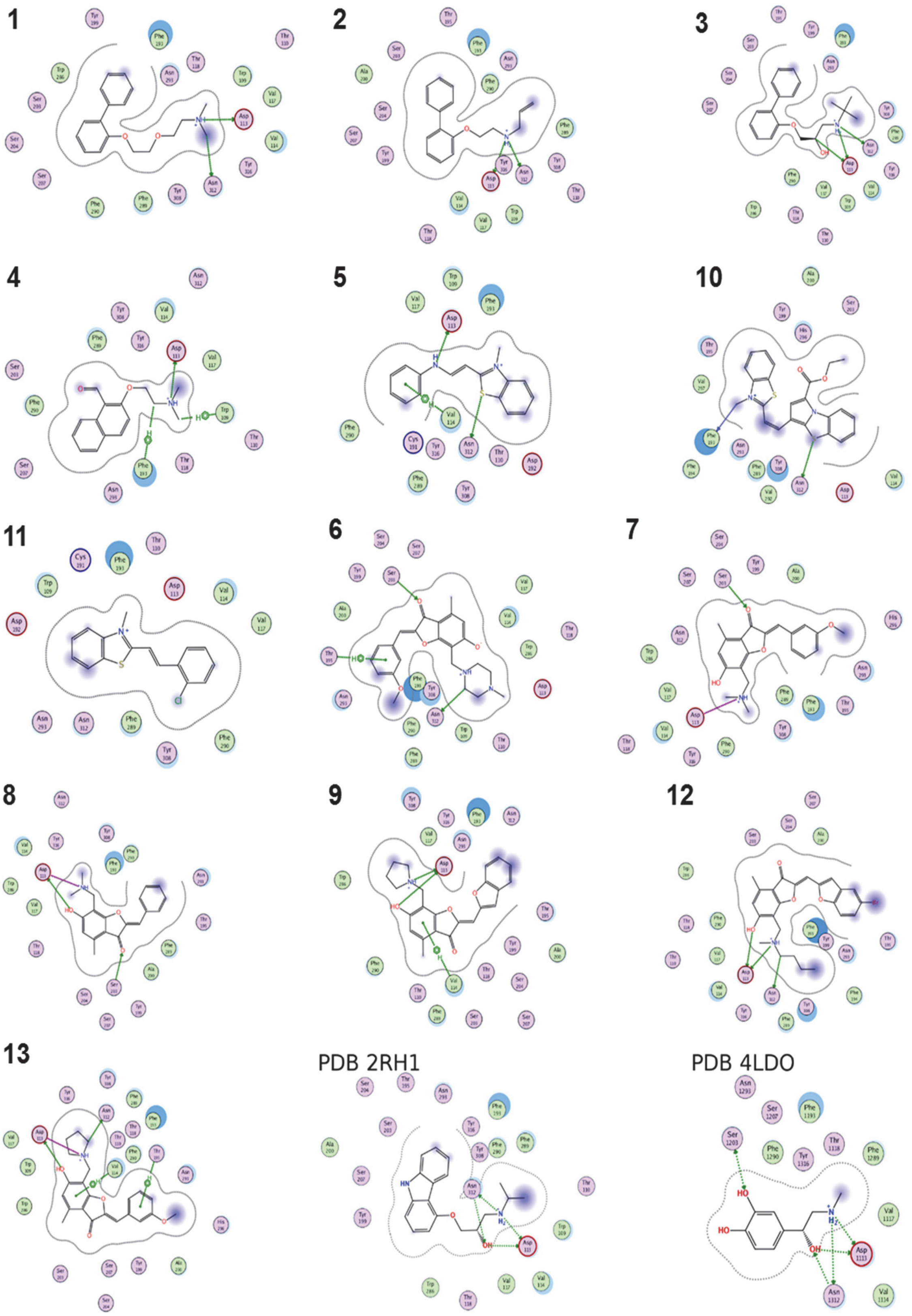


Figure S6: 2D binding mode depictions for all compounds for which binding has been correctly predicted (1,2,3,5,6,10,11,12,13,14,15,16,17). For comparison, the binding modes for Carazolol (PDB 2RH1) and adrenaline (PDB 4LDO) are shown. For new compounds, the depictions have been calculated based on binding mode predicted by docking. Depictions created using the Molecule Operating Environment (MOE). (10)

9. Supplementary charts

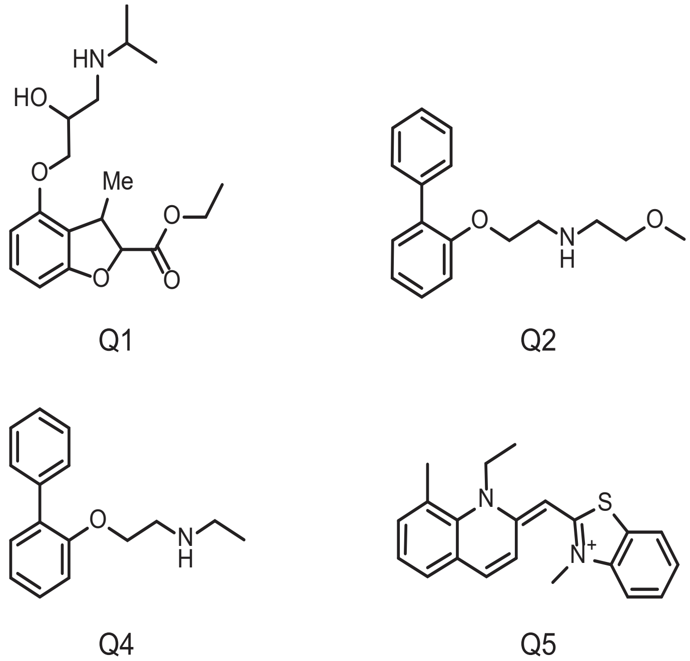


Chart S1: The six query molecules from ref. (5) used for similarity search and the derivation of eight substructures.

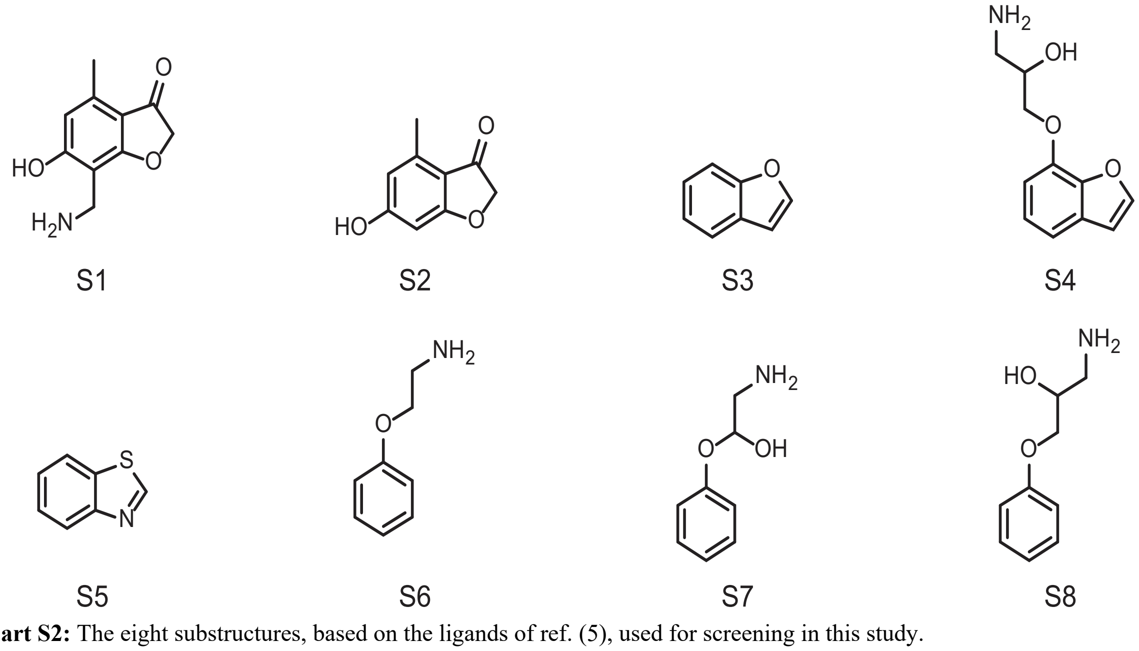


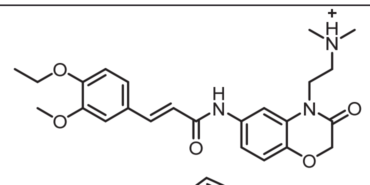
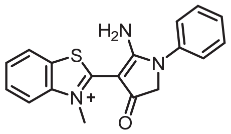
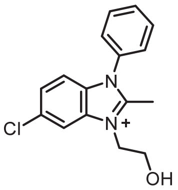
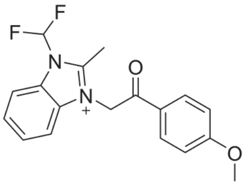
Chart S2: The eight substructures, based on the ligands of ref. (5), used for screening in this study.

10. Supplementary tables

Table S1: Number of molecules resulting from the similarity search with TC ≥ 0.45 for each query molecule of ref. (*2*). The sum reflects the number of molecules after removing duplicates.

Query	Q1	Q2	Q3	Q4	Q5	Q6	Σ
N_{Hits}	1538	2381	946	1310	1053	284	6363

Table S2: Affinity (K_D values) and β_2-selectivity for compounds as measured by [^{3}H](-)CGP 12177 whole cell binding to CHO-β_1 and CHO-β_2 cells. Values are mean ± sem of n separate experiments.

ID	Structure	β_2AR pK_D	n	β_1AR pK_D	n	β_2/β_1[a]
14-4		n.c.	5	n.c.	5	
15-7		n.c.	3	n.c.	3	
16-8		n.c.	5	n.c.	5	
17-9		n.c.	6	n.c.	6	

[a]Selectivity: $\beta_2/\beta_1=K_D(\beta_2)/K_D(\beta_1)$

[n.c.]For ligands with less than 50% inhibition of specific binding, the IC_{50} value could not be determined and thus a K_D value could not be calculated (n.c.)

Table S4: EC_{50} values and % isoprenaline maximum values for cimaterol, 3 and CGP 12177 as agonists and log K_D values for CGP 20712A and ICI 118551 as antagonists of these agonist response in the CHO-β_1 and CHO-β_2 cells respectively, as determined from CRE-SPAP production. Values are mean ± sem of n separate determinations.

Agonist	pEC_{50}		% isoprenaline maximum	n	pK_D		n
β_1AR					CGP 20712A		
cimaterol	8.81	±0.02	104.7 ± 1.9	12	9.21	±0.06	15
3	8.80	±0.06	45.0 ± 2.0	11	8.35	±0.08	21
CGP 12177	8.39	±0.03	86.8 ± 2.8	7	7.47	±0.04	14
β_2AR					ICI 118551		
cimaterol	9.71	±0.02	94.9 ± 1.4	9	9.81	±0.15	5
3	9.94	±0.1	22.0 ± 2.1	8	9.56	±0.06	11

Table S5: Affinity (K_D values) and β_2-selectivity for compounds as measured by a parallel shift inhibition of cimaterol concentration responses in the CRE-SPAP assay in CHO-β_1 and CHO-β_2 cells. Values are mean ± sem of n separate experiments.

ID	β_2AR pK_D		n	β_1AR pK_D		n	β_2/β_1[a]
1[c]	5.63	±0.06	6	4.81	±0.07	6	6.6
2[c]	5.73	±0.06	4	4.85	±0.05	5	7.6
3[b,d]	10.74	±0.03	4	9.19	±0.04	10	35.5
4-5	4.98	±0.05	6	4.62	±0.08	4	2.3
5-6	n.c.		8	n.c.		4	
6-12	5.30	±0.07	4	5.01	±0.06	6	1.9
7-13	5.30	±0.06	4	5.02	±0.07	6	1.9
8-14	5.06	±0.04	4	5.17	±0.12	7	0.78
9-15	n.c.		6	n.c.		6	
10	6.62	±0.12[e]	3	n.c.		3	
11	n.c.		8	n.c.		6	
12-16	n.c.		8	n.c.		6	
13-17	5.58	±0.1	6	4.82	±0.08[e]	5	5.8
14-4	n.c.		4	n.c.		4	
15-7	4.13	±0.06	6	n.c.		4	
16-8	n.c.		8	n.c.		4	
17-9	n.c.		6	n.c.		4	
ICI 118551	9.81	±0.15	5	7.16	±0.07	6	447
CGP 20712A	6.21	±0.04	6	9.21	±0.06	15	0.001

[a] Selectivity: $\beta_2/\beta_1 = K_D(\beta_2)/K_D(\beta_1)$

[b] the partial agonist method of Stephenson 1956 was used to calculate the K_D value for **3**.

[c] US 20090163545

[d] Antiarrythmic pharmaceutical (Bipranol/Berlafenone), Arzneimittel-Forschung **1992**, *42*, 289-291

[e] estimated K_D. Here a shift and a small reduction of the maximum response obtained when incubated with cimaterol rather than an absolute parallel shift was obtained e.g. Figure 2d. The shift was calculated from a parallel response of the lower part of the curve (as per the Gaddam equation) but noted here as an estimated K_D given the slight fall in maximum.

Table S4: SMILES codes, vendor information and ZINC ID for described compounds

ID	SMILES string	Vendor	Vendor ID	ZINC ID
1[b]	CN(C)CCOCCOc1ccccc1-c1ccccc1	Ambinter[a]	Amb8591782	2825338
2[b]	C=CCNCCOc1ccccc1-c1ccccc1	Ambinter[a]	Amb10982638	3001189
3[c]	CC(C)(C)NCC(O)COc1ccccc1-c1ccccc1	Innovapharm[b]	STT-00320296	4353
4-5	CN(C)CCOc1ccc2ccccc2c1C=O	Otava[c]	7020663309	11992987
5-6	C[n+]1c(C=CNc2ccccc2)sc2ccccc21	Enamine[d]	T0504-1129	
6-12	COc1cccc(C=C2Oc3c(c(C)cc(O)c3CN3CCN(C)CC3)C2=O)c1	IBS[f]	STK854129	20573542
7-13	COc1cccc(C=C2Oc3c(c(C)cc(O)c3CN(C)C)C2=O)c1	IBS[f]	Amb807687	6764660
8-14	Cc1cc(O)c(CN(C)C)c2c1C(=O)C(=Cc1ccccc1)O2	IBS[f]	Amb800928	6670218
9-15	Cc1cc([O-])c(C[NH+]2CCCC2)c2c1C(=O)C(=Cc1cc3ccccc3o1)O2	Ambinter[a]	Amb2448714	9575977
10	CCOC(=O)c1cc(C=Cc2sc3ccccc3[n+]2C)c2sc3ccccc3n12	Otava[c]	107910005	15222345
11	C[n+]1c(C=Cc2ccccc2Cl)sc2ccccc21	Otava[c]	107910003	4158946
12-16	CCCC[NH+](C)Cc1c2c(c(C)cc1[O-])C(=O)C(=Cc1cc3cc(Br)ccc3o1)O2	Ambinter[a]	Amb2453954	9531929
13-17	COc1cccc(C=C2Oc3c(c(C)cc(O)c3CN3CCCC3)C2=O)c1	Ambinter[a]	Amb804798	6759304
14-4	CCOc1ccc(C=CC(=O)Nc2ccc3c(c2)N(CCN(C)C)C(=O)CO3)cc1OC	Otava[c]	1082925	12082453
15-7	C[n+]1c2ccccc2sc1C1=C(N)N(c2ccccc2)CC1=O	Ambinter[a]	Amb471924	8394352
16-8	Cc1n(-c2ccccc2)c2ccc(Cl)cc2[n+]1CCO	Ambinter[a]	Amb8495562	3127921
17-9	COc1ccccc1C(=O)C[n+]1c(C)n(C(F)F)c2ccccc21	Timetec[e]	ST51248084	5571431

[a] Ambinter c/o Greenpharma, 3, allée du titane 45100 Orléans, FRANCE

[b] Innovapharm Ltd., 42 Krasnotkatskaya Street, app. 111, Kiev – 02660, UKRAINE

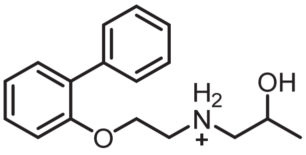
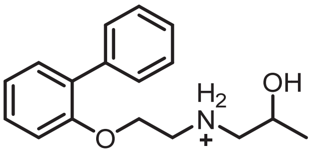
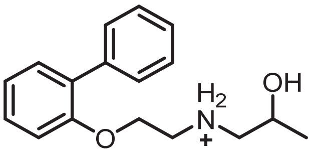
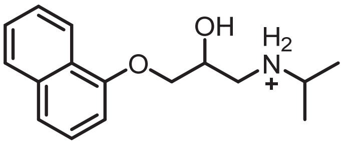
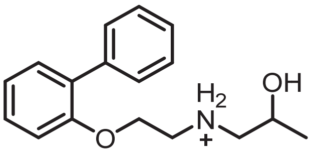
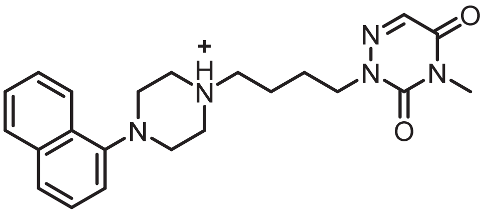
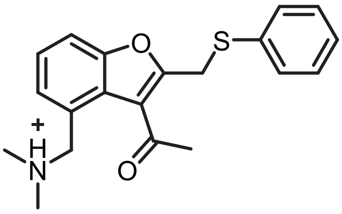
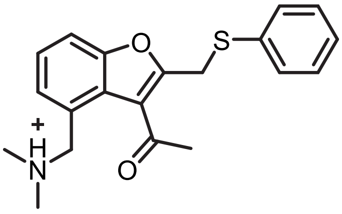
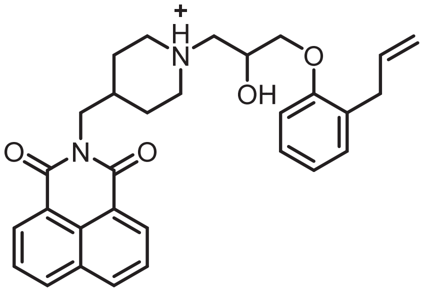
[c] OTAVA Ltd., 400 Applewood Crescent, Unit 100, Vaughan, Ontario, L4K 0C3, CANADA

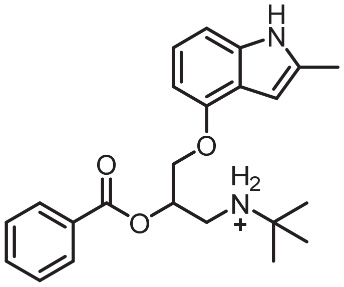
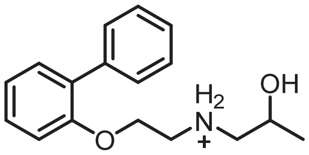
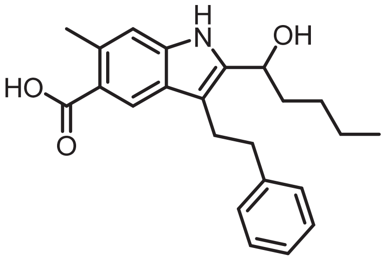
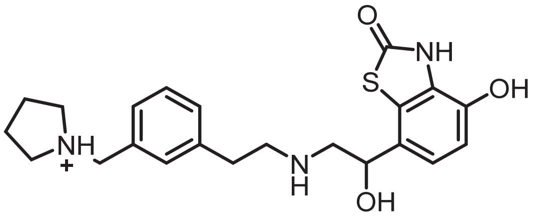
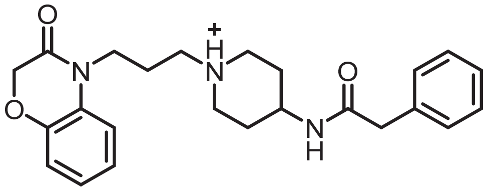
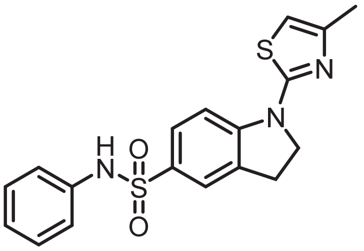
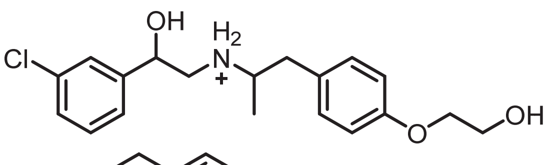
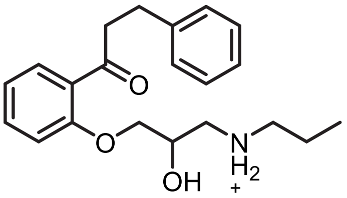
[d] SIA Enamine, Vestienas iela 2 B, V-1035 Riga, LATVIA

[e] TimTec LLC, 301-A Harmony Business Park, Newark, DE 19711, USA

[f] InterBioScreen Ltd., Institutsky Prospect, 7a, 142432 Chernogolovka, RUSSIA

Table S6: Most similar molecules (ChEBML ID and structure) for each compound by Tanimoto ECFP4 similarity at the time of the investigation

ID	ChEMBL ID	Structure	Tanimoto ECFP4
1[a]	CHEMBL1626224		0.5870
2[a]	CHEMBL1626224		0.7270
3[a]	CHEMBL1626224		0.6220
4-5	CHEMBL275742		0.3820
5-6	CHEMBL1626224		0.3260
6-12	CHEMBL2068762		0.3150
7-13	CHEMBL1622248		0.2990
8-14	CHEMBL1622248		0.3960
9-15	CHEMBL1083366		0.2180

10	CHEMBL357995		0.2540
11	CHEMBL1626224		0.2700
12-16	CHEMBL403296		0.2310
13-17	CHEMBL1945294		0.3200
14-4	CHEMBL1242923		0.2710
15-7	CHEMBL433454		0.2900
16-8	CHEMBL41113		0.2890
17-9	CHEMBL631		0.3500

[a] Compounds are annotated in the latest ChEMBL version (ChEMBL 22)

Reference list

1. Irwin, J. J.; Shoichet, B. K. ZINC - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005,** *45* (1), 177-182.
2. Kolb, P.; Rosenbaum, D. M.; Irwin, J. J.; Fung, J. J.; Kobilka, B. K.; Shoichet, B. K. Structure-based discovery of beta2-adrenergic receptor ligands. *Proc. Natl. Acad. Sci. U. S. A.* **2009,** *106* (16), 6843-6848.
3. Baker, J. G. The selectivity of β-adrenoceptor antagonists at the β1, β2 and β3 adrenoceptors. *Br. J. Pharmacol.* **2005,** *144,* 317-322.
4. Baker, J. G.; Proudman, R. G. W.; Hill, S. J. Identification of key residues in transmembrane 4 responsible for the secondary, low affinity conformation of the human β1-adrenoceptor. *Mol. Pharmacol.* **2014,** *85,* 811-829.
5. Pak, M. D.; Fishman, P. H. Anomalous behaviour of CGP 12177A on β1-adrenergic receptors. *J. Recept. Signal Transduction Res.* **1996,** *16,* 1-23.
6. Konkar, A. A.; Zhengxian, Z.; Granneman, J. G. Aryloxypropanolamine and catecholamine ligand interactions with the β1-adrenergic receptor: evidence for interaction with distinct conformations of β1-adrenergic receptors. *J. Pharmacol. Exp. Ther.* **2000,** *294,* 923-932.
7. Baker, J. G. Site of Action of β-Ligands at the Human β1-Adrenoceptor. *J. Pharmacol. Exp. Ther.* **2005,** *313* (3), 1163-1171.
8. Kaumann, A. J.; Molenaar, P. The low-affinity site of the β1-adrenoceptor and its relevance to cardiovascular pharmacology. *Pharmacol. Ther.* **2008,** *118,* 303-336.
9. Baker, J. G.; Hall, I. P.; Hill, S. J. Agonist and inverse agonist actions of "β-blockers" at the human β2-adrenoceptor provide evidence for agonist-directed signalling. *Mol. Pharmacol.* **2003,** *64,* 1357-1369.
10. Chemical Computing Group, Inc. Molecular Operating Environment (MOE) 2015.10. *1010 Sheerbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7,* **2015**.